VISIT…

LANZAROTE

Caliente.COM

УДК 519.6
ББК 22.19
Ф 79

Формалев В. Ф., Ревизников Д. Л. **Численные методы.** — М.: ФИЗМАТЛИТ, 2004. — 400 с. — ISBN 5-9221-0479-9.

В учебнике представлены основные численные методы решения задач алгебры и анализа, теории приближений и оптимизации, задач для обыкновенных дифференциальных уравнений и уравнений математической физики. Систематически изложены методы конечных разностей, конечных и граничных элементов, методы исследования аппроксимации, устойчивости, сходимости, оценок погрешности. Каждый метод иллюстрируется подробно разобранным примером, даны упражнения для самостоятельной проработки.

Для студентов и аспирантов технических университетов, специализирующихся в области теплотехники, прикладной механики и прикладной математики. Книга ориентирована на двухсеместровый курс обучения.

ISBN 5-9221-0479-9

© ФИЗМАТЛИТ, 2004
© В. Ф. Формалев, Д. Л. Ревизников, 2004

ПРЕДИСЛОВИЕ

Данный учебник написан на основе двухсеместрового курса лекций, читавшегося авторами на протяжении пятнадцати лет студентам математических и механических специальностей Московского авиационного института (государственного технического университета) (МАИ). Он состоит из двух частей. Первая часть учебника содержит следующие разделы: элементы теории погрешностей, численные методы алгебры (как линейной, так и общей), теорию приближений, численные методы решения задач для обыкновенных дифференциальных уравнений и численные методы оптимизации.

Вторая часть включает в себя численные методы решения задач математической физики, исследование аппроксимации, устойчивости, сходимости и консервативности конечно-разностных, конечно-элементных и гранично-элементных методов. Кроме этого, во вторую часть учебника вошли разделы по методу конечных разностей решения многомерных задач математической физики, среди которых такие методы, как методы расщепления, установления, прямых, характеристик, метод С. К. Годунова, а также методы конечных и граничных элементов с анализом погрешностей.

Отбирая материал для учебника, авторы не загромождали его изложением всего множества существующих алгоритмов, а ограничились описанием наиболее широко используемых на практике и популярных методов, а также изложением новых эффективных методов.

Вместе с тем изложенные методы вычислений описаны на таком конструктивном уровне, с использованием решенных примеров и упражнений, что читателю нет необходимости при освоении материала обращаться к другим источникам по вычислительной математике. Правда, он должен обладать знанием основ линейной алгебры, дифференциального и интегрального исчислений, обыкновенных дифференциальных уравнений, а для освоения второй части — основами теории уравнений в частных производных.

Книга состоит из девяти глав, первые пять из которых составляют первую часть и содержат материал стандартного односеместрового курса численных методов, причем поскольку численные методы оптимизации читаются, как правило, в отдельном курсе, то в содержание главы 5 включены только алгоритмы безусловной минимизации.

В существующих учебниках отсутствует систематическое изложение метода конечных элементов. Поэтому в данном учебнике, наряду с систематическим изложением метода конечных разностей в задачах для обыкновенных дифференциальных уравнений (ОДУ) и уравнений математической физики, подробно изложены методы конечных и граничных элементов на основе метода взвешенных невязок Галеркина, а также метод конечных элементов на основе вариационного принципа.

Кроме этого, в раздел «Методы расщепления численного решения многомерных задач математической физики» включены разработанные авторами новые, экономичные, абсолютно устойчивые методы численного решения многомерных параболических задач, содержащих смешанные дифференциальные операторы. Приведено доказательство теорем об аппроксимации и абсолютной устойчивости.

В отличие от первой части, где для усвоения каждого метода разбирается соответствующий пример, вторая часть изложена на основе конкретных задач, обобщающих материал рассматриваемого раздела.

Таким образом, с позиции содержания материала учебник носит универсальный характер. В этой связи первую часть учебника можно рекомендовать студентам, обучающимся по *экономическим и техническим специальностям* с односеместровым курсом «Численные методы» или «Численные методы и алгоритмы». Студентам, обучающимся по специальностям «Математическая экономика», «Прикладная математика», «Прикладная механика» и другим специальностям с углубленным изучением математических дисциплин, рекомендуется, наряду с первой частью, и вторая часть учебника, соответствующая двухсеместровому курсу численных методов.

Кроме студентов, обучающихся по упомянутым выше специальностям, учебник полезен аспирантам, инженерам, научным работникам, а также преподавателям вычислительной матема-

тики, которые, кроме всего прочего, могут использовать вторую часть для чтения специальных курсов.

Значительное влияние на содержание учебника оказал чл.-корр. РАН, заведующий кафедрой «Вычислительная математика и программирование» МАИ, профессор У. Г. Пирумов. Авторы выражают ему свою глубокую благодарность.

Авторы глубоко признательны и аспирантам С. В. Миканеву, С. А. Колеснику, Т. А. Тихоновой, оказавшим неоценимую помощь при компьютерном наборе рукописи учебника.

ВВЕДЕНИЕ

В настоящее время появилось значительное число различных программных продуктов (MathCad, Mathlab и т.д.), с помощью которых, задавая только входные данные и *не вникая в сущность алгоритмов*, можно решить значительное число задач, на обучение решению которых и направлен данный учебник.

Безусловно, умение пользоваться этими программными продуктами существенно сокращает время и ресурсы по решению ряда важных задач. Вместе с этим бездумное использование упомянутых выше программ без тщательного анализа метода, с помощью которого решается задача, таит в себе следующие опасности.

Во-первых, все методы имеют ограничения по входным параметрам (например, по размерам матриц при решении систем линейных алгебраических уравнений), и попытка решить задачу с входными параметрами за пределами этих ограничений приводит к неудаче.

Во-вторых, сами методы, в основном численные, имеют существенные ограничения по применению.

В-третьих, незнание метода, с помощью которого решалась конкретная задача в програмном продукте, приводит к ситуации, когда трудно проанализировать качество решения (например, погрешность, скорость сходимости итерационных процессов, устойчивость и другие важнейшие характеристики численных методов).

В-четвертых, стандартные программные продукты значительно ограничены количеством решаемых задач, среди которых в основном линейные задачи. Вне сферы их применения остается большинство задач, связанных с уравнениями математической физики и др.

Поэтому при численном решении задач (а именно с помощью численных методов решается подавляющее число современных задач) вычислитель должен четко представлять каждый из следующих этапов: построение адекватной математической модели, выбор метода численного решения, разработку алгоритма, составление программы, формальную, а затем и фактическую

отладку программы, корректировку и исправление всех этапов, начиная с математической модели, на основе анализа тестовых результатов.

Как видно из перечисления этапов решения задач, путь от постановки до получения результатов не краток. В этой связи следует заметить, что если неопытный вычислитель после осуществления первых четырех этапов считает, что задача решена, то опытный знает, что первый находится в начале сложного пути с неожиданными результатами и поворотами.

Учебник состоит из девяти глав, неравнозначных по объему.

Самая краткая *первая глава* содержит элементы теории погрешностей. В ней авторы обращают внимание на тот факт, что абсолютная погрешность не превышает точности вычислений. С помощью введенного неравенства легко проясняется существо фразы: «Решить задачу с точностью ε», хотя всюду имеем дело с оценкой погрешностей.

Вторая глава посвящена численным методам решения задач алгебры (как линейной, так и общей). Здесь внимание обращается на следующие вопросы, связанные с применением итерационных методов:

1. Сходится ли итерационная последовательность, т. е. существует ли предел итерационной последовательности при стремлении количества итераций в бесконечность?

2. Если такое предельное значение существует, то является ли оно решением задачи?

3. Если получен положительный ответ на первые два вопроса, то, останавливая итерационный процесс на какой-либо итерации, необходимо оценить сверху погрешность решения итерационным методом по сравнению с точным (неизвестным, но существующим) решением. Если при этом задана точность вычислений, то из неравенства «погрешность не превышает точность» можно оценить нижнюю границу числа итераций для достижения заданной точности.

Третья глава содержит основы теории приближений и включает следующие разделы: интерполяцию, включая и сплайн-интерполяцию, аппроксимацию с помощью метода наименьших квадратов, численное дифференцирование и численное интегрирование. При этом обращается внимание на то, что метод численного дифференцирования с помощью отношения конечных разностей с оценкой погрешности метода является осно-

вой конечно-разностного решения задач как для обыкновенных дифференциальных уравнений, так и для уравнений в частных производных.

Четвертая глава посвящена численным методам решения задач для обыкновенных дифференциальных уравнений (задач Коши и краевых задач). При этом известные и хорошо зарекомендовавшие себя методы численного решения задач Коши, такие как методы Эйлера, Эйлера–Коши, Рунге–Кутта, сформулированы соответственно на основе квадратурных формул прямоугольников, трапеций, Симпсона, что позволяет использовать для оценки погрешности соответствующие оценки для квадратурных формул.

Конечно-разностный метод решения краевых задач для ОДУ с граничными условиями, содержащими производные, описан так, что порядок аппроксимации дифференциального уравнения сохранен при аппроксимации краевых условий. Это достигается допущением о том, что гладкость решений на границах совпадает со старшей производной, входящей в дифференциальное уравнение и дальнейшим использованием дифференциального уравнения при аппроксимации краевых условий. Ниже, в шестой главе будет показано, что такой подход не просто выравнивает порядок аппроксимации краевых условий с порядком аппроксимации дифференциального уравнения, но и делает конечно-разностную схему *консервативной*, сохраняющей фундаментальные законы, на основе которых выведены дифференциальное уравнение и краевые условия.

В *пятую главу* включен материал по численным методам, безусловной минимизации функций одной и многих переменных, причем, как отмечалось выше, более сложные аспекты численных методов математического программирования включаются в отдельные курсы.

Шестая глава посвящена методам решения задач для уравнений математической физики. Здесь, кроме конечно-разностных схем в задачах для уравнений гиперболического, параболического и эллиптического типов и алгоритмов их решения, сделан акцент на строгое определение понятий аппроксимации, порядка аппроксимации, устойчивости, сходимости, порядка сходимости (порядка точности), консервативности, а также на описание различных методов исследования устойчивости и нахождения порядка аппроксимации. Проанализированы неявно-

явные конечно-разностные схемы, как следствие двусторонних методов, с исследованием аппроксимации и устойчивости для схем типа Кранка–Николсона. Идеи, заложенные в этой схеме, активно используются в методах расщепления для многомерных задач математической физики. Здесь показано, что конечно-разностная аппроксимация производных первого порядка в краевых условиях с помощью отношения односторонних разностей без использования дифференциального уравнения приводит, с одной стороны, к схемам пониженного порядка, а с другой — к схемам, не обладающим свойством консервативности. То есть приведен пример неконсервативной схемы, пользоваться которой не рекомендуется, так как численно решается задача, отличная от дифференциальной задачи.

В главе на основе трудов А. А. Самарского подробно описан энергетический метод исследования устойчивости конечно-разностных схем, в котором энергетическое тождество приводит к достаточным условиям устойчивости вследствие выполнения принципа максимума. Метод гармонического анализа, также подробно изложенный, дает только необходимые условия устойчивости.

В седьмой главе приведены численные методы решения многомерных задач математической физики, такие как методы расщепления, метод характеристик для квазилинейных гиперболических систем, метод прямых, метод сквозного счета С. К. Годунова. Наряду с широко известными методами переменных направлений Писмена–Рэчфорда и дробных шагов Н.Н. Яненко, рассмотрены экономичные абсолютно устойчивые методы переменных направлений с экстраполяцией и полного расщепления, принадлежащие авторам. В отличие от существующих методов, предложенные методы применимы к задачам, содержащим смешанные производные и любую размерность по пространственным переменным с сохранением при этом порядка аппроксимации и *абсолютной устойчивости*, что обосновывается доказательством соответствующих теорем. Для всех методов расщепления исследованы порядок аппроксимации и устойчивость.

Восьмая и *девятая главы* посвящены систематическому изложению методов конечных и граничных элементов соответственно. Основное внимание уделено методу взвешенных невязок Галеркина, поскольку вариационные методы в методе конечных

элементов (МКЭ) существенно ограничивают круг решаемых задач, так как не для всякой задачи математической физики можно построить вариационный функционал. В главе показано, что вариационный функционал с использованием метода Релея–Ритца можно всегда построить для дифференциальных уравнений, содержащих симметрический дифференциальный оператор. Изложен метод нахождения погрешности конечно-элементного решения в классе функций, принадлежащих пространству Соболева W_2^1. В этот класс, как известно, входят кусочно-линейные базисные функции, на основе которых строится решение.

Метод граничных элементов изложен для стационарных задач математической физики в многосвязных пространственных областях, хотя для нестационарных задач можно поступить так же, как в методе конечных элементов, где дифференциальный оператор по времени аппроксимирован с помощью отношения конечных разностей, а пространственные — с помощью метода конечных элементов.

Часть I

ЧИСЛЕННЫЕ МЕТОДЫ АЛГЕБРЫ И АНАЛИЗА

В первую часть включены следующие разделы стандартного односеместрового курса «Численные методы»: 1) элементы теории погрешностей; 2) численные методы алгебры; 3) теория приближений; 4) численные методы решения задач для обыкновенных дифференциальных уравнений; 5) численные методы оптимизации.

ГЛАВА I

ЭЛЕМЕНТЫ ТЕОРИИ ПОГРЕШНОСТЕЙ

Программа

Погрешности, их источники, устранимые и неустранимые погрешности. Абсолютная и предельно абсолютная погрешности. Абсолютные погрешности выражений. Значащие и верные цифры. Точность, соотношение между погрешностью и точностью. Округление чисел.

Погрешность решения задачи обусловливается следующими причинами.

1. Математическое описание и исходные данные являются неточными.

2. Применяемые методы являются чаще всего приближенными, мало того, решение не может быть получено за конечное число арифметических операций.

3. В процессе вычисления проводятся округления. В соответствии с этим погрешности называют:

1) *неустранимыми погрешностями*;

2) *погрешностями метода*;

3) *вычислительными погрешностями*.

Пусть a — приближенное число для точного числа A. Погрешностью Δa приближенного числа a называют разность

$$\Delta a = A - a.$$

Абсолютной погрешностью $\Delta(a)$ приближенного числа a называют абсолютную величину погрешности

$$\Delta(a) = |A - a|,$$

позволяющую отвлечься от знака погрешности.

Поскольку в большинстве случаев точное значение числа A неизвестно и, следовательно, невозможно вычислить абсолютную погрешность, вводят понятие *предельной абсолютной погрешности* Δ_a приближенного числа a, удовлетворяющее следующему соотношению:

$$\Delta(a) = |A - a| \leqslant \Delta_a,$$

или

$$a - \Delta_a \leqslant A \leqslant a + \Delta_a.$$

При вычислении абсолютных (предельных абсолютных) погрешностей выражений удобно использовать формулы дифференцирования, заменяя дифференциалы независимых переменных абсолютными (предельными абсолютными) погрешностями.

Пример 1.1.

$$\Delta(a + b) = \Delta(a) + \Delta(b) \leqslant \Delta_a + \Delta_b;$$

$$\Delta(a - b) = \Delta(a) - \Delta(b) \leqslant \Delta_a + \Delta_b.$$

Пример 1.2.

$$\Delta(a \cdot b) = a\Delta(b) + b\Delta(a) \leqslant a\Delta_b + b\Delta_a.$$

Пример 1.3.

$$\Delta\left(\frac{a}{b}\right) = \frac{\Delta(a)b - a\Delta(b)}{b^2} \leqslant \frac{a\Delta_b + b\Delta_a}{b^2}.$$

Пример 1.4. Вычислить $\Delta\left(\left(\frac{a + b \cdot c}{f}\right)^2\right)$.

Решение. Пусть $du(a,b,c,f) \approx \Delta\left(\left(\frac{a+b\cdot c}{f}\right)^2\right)$. Тогда

$$\Delta(u) \approx \frac{\partial u}{\partial a}\Delta(a) + \frac{\partial u}{\partial b}\Delta(b) + \frac{\partial u}{\partial c}\Delta(c) + \frac{\partial u}{\partial f}\Delta(f) =$$

$$= 2\left(\frac{a+b\cdot c}{f}\right)\left[\frac{f}{f^2}\Delta(a) + \frac{fc}{f^2}\Delta(b) + \frac{fb}{f^2}\Delta(c) + \frac{-(a+bc)}{f^2}\Delta(f)\right] \leqslant$$

$$\leqslant 2\left(\frac{a+b\cdot c}{f}\right)\left(\frac{1}{f}\Delta_a + \frac{c}{f}\Delta_b + \frac{b}{f}\Delta_c + \frac{a+bc}{f^2}\Delta_f\right).$$

Значащими цифрами приближенного числа a называются все цифры в его записи, начиная с первой ненулевой слева.

Пример 1.5. $a = 0{,}0\underline{2087}$, $a = 0{,}0\underline{208700}$ (значащие цифры подчеркнуты).

Если приближенное число a имеет *n значащих цифр* , то за предельную абсолютную погрешность числа a принимают половину единицы разряда, выражаемого n-й значащей цифрой, считая слева направо.

Пример 1.6. Пусть даны величины a с предельными абсолютными погрешностями Δ_a:

a	$-2{,}17$	$3{,}141$	$0{,}012$	1795
Δ_a	$0{,}005$	$0{,}0005$	$0{,}0005$	$0{,}5$

Тогда, например, для числа $a = -2{,}17:\ -2{,}175 \leqslant A \leqslant -2{,}165$.

Относительной погрешностью $\delta(a)$ приближенного числа a называется отношение в долях единицы:

$$\delta(a) = \frac{|A-a|}{|a|} = \frac{\Delta(a)}{|a|}.$$

Предельной относительной погрешностью δ_a приближенного числа a называют число, не меньшее относительной погрешности этого числа:

$$\delta(a) = \frac{|A-a|}{|a|} \leqslant \delta_a,$$

или

$$a - |a|\,\delta_a \leqslant A \leqslant\; a + |a|\,\delta_a.$$

Ясно, что

$$\delta_a = \frac{\Delta_a}{|a|}.$$

Значащую цифру приближенного числа a называют *верной*, если абсолютная погрешность этого числа не превышает половины единицы разряда, соответствующего этой цифре (не превышает предельной абсолютной погрешности числа).

Пример 1.7. Для приближенного числа $a = 0{,}02087$ с абсолютной погрешностью $\Delta(a) = 0{,}4 \cdot 10^{-5}$ определить верные значащие цифры.

Решение. Поскольку данная погрешность числа не превышает предельной абсолютной погрешности этого числа, равной $\Delta_a = 0{,}5 \cdot 10^{-5}$, т. е. $\Delta(a) = 0{,}4 \cdot 10^{-5} < \Delta_a = 0{,}5 \cdot 10^{-5}$, то верными будут значащие цифры 7, 8, 0, 2, т. е. $a = 0{,}0\underline{2087}$ (верные цифры подчеркнуты).

Вычислить приближенное число a *с точностью до* $\varepsilon = 10^{-n}$ означает необходимость сохранить верной значащую цифру, стоящую в n-м разряде после запятой.

Пример 1.8. Вычислить $\sqrt{2}$ с точностью $\varepsilon = 10^{-3}$.

Решение. $\sqrt{2} = \underline{1{,}414}2$; третья цифра после запятой является верной, т. к. $\Delta(\sqrt{2}) = |1{,}4142 - 1{,}414| = 0{,}0002 < 0{,}0005 = \Delta a < \varepsilon = 10^{-3}$. Следовательно, все подчеркнутые цифры являются верными.

Таким образом, из определения абсолютной погрешности приближенного числа a и точности его вычисления вытекает очевидная связь:

$$\Delta(a) \leqslant \Delta_a < \varepsilon,$$

т. е. *абсолютная погрешность и предельная абсолютная погрешность не превышают точности.*

Относительная погрешность суммы, разности, произведения, частного не превышает суммы относительных погрешностей операндов: $\delta(a \pm b) \leqslant \delta(a) + \delta(b)$; $\delta(a \cdot b) \leqslant \delta(a) + \delta(b)$; $\delta\,(a/b) \leqslant \delta(a) + \delta(b)$. Относительная погрешность степени не

превышает произведения показателя степени на относительную погрешность основания: $\delta(a^m) \leqslant m\delta(a)$; $\delta(\sqrt[m]{a}) \leqslant \delta(a)/m$.

При округлении приближенного числа a до n-й значащей цифры необходимо к цифре $(n+1)$-го разряда прибавить цифру 5; если полученное число больше или равно 10, то к цифре n-го разряда добавляется единица, а разряды начиная с $(n+1)$-го отбрасываются; в противном случае разряды начиная с $(n+1)$-го отбрасываются без прибавления единицы к n-му разряду.

Пример 1.9. Округлить число $\pi = 3{,}141592$: а) до третьей значащей цифры; б) до четвертой значащей цифры.

Решение. а) 3,14, так как $1{,}592 + 5 = 6{,}592 < 10$; б) 3,142, так как $5{,}92 + 5 = 10{,}92 > 10$.

ГЛАВА II

ЧИСЛЕННЫЕ МЕТОДЫ АЛГЕБРЫ

Программа

Прямые методы решения систем линейных алгебраических уравнений (СЛАУ). Метод Гаусса, его применение для обращения и вычисления определителей матриц. Метод прогонки и циклической прогонки, их обоснование. Матричная прогонка. Нормы векторов и матриц, их согласование. Итерационные методы решения СЛАУ. Метод простых итераций и метод Зейделя с обоснованием сходимости. Методы решения нелинейных уравнений. Методы отделения и уточнения корней, геометрический смысл, сходимость, погрешность методов: половинного деления, Ньютона, секущих, простых итераций. Скорость сходимости, методы ускорения сходимости. Методы простых итераций, Зейделя и Ньютона решения систем нелинейных уравнений. Численные методы решения задач на собственные значения и собственные векторы матриц. Метод вращения Якоби, степенной метод.

В главе «Численные методы алгебры» рассматриваются численные методы решения систем линейных алгебраических уравнений (СЛАУ), численные методы решения нелинейных уравнений и систем нелинейных уравнений, численные методы решения задач на собственные значения и собственные векторы матриц.

Среди численных методов алгебры существуют прямые методы, в которых решение получается за конечное число операций, и итерационные методы, в которых результат получается за бесконечное число операций, но прерывая этот процесс на какой-либо итерации необходимо сделать следующее:

1) доказать сходимость итерационной последовательности;

2) если итерационная последовательность сходится, необходимо определить, является ли предельное значение решением задачи;

3) при утвердительном ответе на первые два вопроса и остановке процесса на какой-либо итерации необходимо уметь оценить погрешность итерационного значения по сравнению с точным решением, которое неизвестно.

§ 2.1. Численные методы решения СЛАУ

Из прямых методов решения СЛАУ рассмотрим методы Гаусса и прогонки [1,2].

2.1.1. Метод Гаусса. В методе Гаусса матрица СЛАУ с помощью элементарных алгебраических операций преобразуется в верхнюю (нижнюю) треугольную матрицу, получающуюся в результате прямого хода. В обратном ходе определяются неизвестные.

Пусть дана СЛАУ

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{cases}$$

Запишем расширенную матрицу системы с контрольными суммами:

Ведущая строка ↓

x_1	$x_2\ x_3 \ldots x_n$	b	k_Σ
a_{11}	$a_{12}\, a_{13} \ldots a_{1n}$	b_1	$\Sigma_1 = b_1 + \sum\limits_{j=1}^{n} a_{1j}$
a_{21}	$a_{22}\, a_{23} \ldots a_{2n}$	b_2	$\Sigma_2 = b_2 + \sum\limits_{j=1}^{n} a_{2j}$
a_{31}	$a_{32}\, a_{33} \ldots a_{3n}$	b_3	$\Sigma_3 = b_3 + \sum\limits_{j=1}^{n} a_{3j}$
$\ldots$	$\ldots\ldots\ldots\ldots$	$\ldots$	$\ldots\ldots\ldots\ldots$
a_{n1}	$a_{n2} a_{n3} \ldots a_{nn}$	b_n	$\Sigma_n = b_n + \sum\limits_{j=1}^{n} a_{nj}$

$\left(-\frac{a_{21}}{a_{11}}\right); \left(-\frac{a_{31}}{a_{11}}\right); \ldots; \left(-\frac{a_{n1}}{a_{11}}\right)$

$\xrightarrow{\text{1-й шаг}}$

← *Ведущий столбец*

На первом шаге алгоритма Гаусса выберем диагональный элемент $a_{11} \neq 0$ (если он равен 0, то первую строку переставляем с какой-либо нижележащей строкой) и объявляем его *ведущим*, а соответствующую строку и столбец, на пересечении которых он стоит, — *ведущими*. Обнулим элементы ведущего столбца,

находящегося под ведущим элементом. Для этого сформируем числа $\left(-\frac{a_{21}}{a_{11}}\right)$; $\left(-\frac{a_{31}}{a_{11}}\right)$; ... ; $\left(-\frac{a_{n1}}{a_{11}}\right)$ и выпишем их около ведущей строки. Умножая ведущую строку на число $\left(-\frac{a_{21}}{a_{11}}\right)$, складывая со второй и ставя результат на место второй строки, получим вместо элемента a_{21} нуль, а вместо элементов a_{2j}, $j = \overline{2,n}$, b_2 и Σ_2 соответственно элементы $a_{2j}^1 = a_{2j} + a_{1j}\left(-\frac{a_{21}}{a_{11}}\right)$, $j = \overline{2,n}$, $b_2^1 = b_2 + b_1\left(-\frac{a_{21}}{a_{11}}\right)$, $\Sigma_2^1 = \Sigma_2 + \Sigma_1\left(-\frac{a_{21}}{a_{11}}\right)$. И так далее. Умножая ведущую строку на число $\left(-\frac{a_{n1}}{a_{11}}\right)$, складывая с n-ой строкой и ставя результат на место n-ой строки, получим вместо элемента a_{n1} нуль, а остальные элементы этой строки будут иметь вид: $a_{nj}^1 = a_{nj} + a_{1j}\left(-\frac{a_{n1}}{a_{11}}\right)$, $b_n^1 = b_n + b_1\left(-\frac{a_{n1}}{a_{11}}\right)$, $\Sigma_n^1 = \Sigma_n + \Sigma_1\left(-\frac{a_{n1}}{a_{11}}\right)$. Сохраняя ведущую строку неизменной, получим в результате 1-го шага алгоритма Гаусса следующую матрицу (при этом сумма преобразованных элементов какой-либо строки и правой части должна быть равна преобразованной по тому же алгоритму контрольной сумме; если это не так, то в соответствующей строке сделана ошибка, которую необходимо устранить):

x_1	x_2	$x_3 \ldots x_n$	b	k_Σ
a_{11}	a_{12}	$a_{13} \ldots a_{1n}$	b_1	Σ_1
0	a_{22}^1	$a_{23}^1 \ldots a_{2n}^1$	b_2^1	Σ_2^1
0	a_{32}^1	$a_{33}^1 \ldots a_{3n}^1$	b_3^1	Σ_3^1
...	...	....	...	...
0	a_{n2}^1	$a_{n3}^1 \ldots a_{nn}^1$	b_n^1	Σ_n^1

Ведущая строка

$\left(-\frac{a_{32}^1}{a_{22}^1}\right); \ldots ; \left(-\frac{a_{n2}^1}{a_{22}^1}\right)$

2-й шаг

← *Ведущий столбец*

На втором шаге алгоритма Гаусса в качестве ведущего элемента выбирается элемент $a_{22}^1 \neq 0$ (если он равен нулю, то вто-

рую строку меняем местами с *нижележащей* строкой). Формируются числа $\left(-\dfrac{a_{32}^1}{a_{22}^1}\right)$; ... ; $\left(-\dfrac{a_{n2}^1}{a_{22}^1}\right)$, которые ставятся около ведущей строки. Умножая ведущую строку на число $\left(-\dfrac{a_{32}^1}{a_{22}^1}\right)$, складывая с третьей строкой и ставя результат на место третьей строки, получим вместо элемента a_{32}^1 нуль, а вместо элементов a_{3j}^1, $j = \overline{3, n}$, b_3^1, Σ_3^1 — элементы $a_{3j}^2 = a_{3j}^1 + a_{2j}^1\left(-\dfrac{a_{32}^1}{a_{22}^1}\right)$, $j = \overline{3, n}$, $b_3^2 = b_3^1 + b_2^1\left(-\dfrac{a_{32}^1}{a_{22}^1}\right)$, $\Sigma_3^2 = \Sigma_3^1 + \Sigma_2^1\left(-\dfrac{a_{32}^1}{a_{22}^1}\right)$. И так далее. Умножая ведущую строку на число $\left(-\dfrac{a_{n2}^1}{a_{22}^1}\right)$, складывая результат с n-ой строкой и ставя полученную сумму на место n-ой строки, получим вместо элемента a_{n2}^1 нуль, а вместо элементов a_{nj}^1, b_n^1, Σ_n^1 — элементы $a_{nj}^2 = a_{nj}^1 + a_{2j}^1\left(-\dfrac{a_{n2}^1}{a_{22}^1}\right)$, $j = \overline{3, n}$, $b_n^2 = b_n^1 + b_2^1\left(-\dfrac{a_{n2}^1}{a_{22}^1}\right)$, $\Sigma_n^2 = \Sigma_n^1 + \Sigma_2^1\left(-\dfrac{a_{n2}^1}{a_{22}^1}\right)$. Сохраняя 1-ю и 2-ю строки матрицы неизменными, получим в результате второго шага алгоритма Гаусса следующую матрицу (при этом сумма преобразованных элементов какой-либо строки и правой части должна быть равна преобразованной контрольной сумме; если это не так, то в соответствующей строке сделана ошибка):

$$
\begin{array}{c}
\begin{array}{ccccccc} x_1 & x_2 & x_3 & \ldots & x_n & b & k_\Sigma \end{array} \\
\left[\begin{array}{ccc|c|c|c}
a_{11} & a_{12} & a_{13} & \ldots a_{1n} & b_1 & \Sigma_1 \\
0 & a_{22}^1 & a_{23}^1 & \ldots a_{2n}^1 & b_2^1 & \Sigma_2^1 \\
\hline
0 & 0 & a_{33}^2 & \ldots a_{3n}^2 & b_3^2 & \Sigma_3^2 \\
\hline
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & a_{n3}^2 & \ldots a_{nn}^2 & b_n^2 & \Sigma_n^2
\end{array}\right]
\end{array}
\xrightarrow[\text{3-й шаг}]{} \ldots \xrightarrow[(n-1)\text{-й шаг}]{}
$$

После $(n-1)$-го шага алгоритма Гаусса получаем следующую расширенную матрицу с контрольными суммами, содержа-

щую верхнюю треугольную матрицу СЛАУ:

$$
\begin{array}{c}
\underline{\quad x_1 \quad x_2 \quad x_3 \ \ldots \ x_n \qquad b \qquad k_\Sigma} \\
\left[\begin{array}{ccccc|c|c}
a_{11} & a_{12} & a_{13} & \ldots & a_{1n} & b_1 & \Sigma_1 \\
0 & a^1_{22} & a^1_{23} & \ldots & a^1_{2n} & b^1_2 & \Sigma^1_2 \\
0 & 0 & a^2_{33} & \ldots & a^2_{3n} & b^2_3 & \Sigma^2_3 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & a^{n-1}_{nn} & b^{n-1}_n & \Sigma^{n-1}_n
\end{array}\right]
\end{array}
$$

Прямой ход алгоритма Гаусса завершен.

В обратном ходе алгоритма Гаусса из последнего уравнения сразу определяется x_n, из предпоследнего — x_{n-1} и т. д. Из первого уравнения определяется x_1:

$$
\begin{cases}
a^{n-1}_{nn}x_n = b^{n-1}_n & \Rightarrow x_n, \\
a^{n-2}_{n-1n-1}x_{n-1} + a^{n-2}_{n-1n}x_n = b^{n-2}_{n-1} & \Rightarrow x_{n-1}, \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a_{11}x_1 + \ldots + a_{1n}x_n = b_1 & \Rightarrow x_1.
\end{cases}
$$

Замечание 1. Если элементы какой-либо строки матрицы системы в результате преобразований стали равными нулю, а правая часть не равна нулю, то СЛАУ несовместна, поскольку не выполняются условия теоремы Кронекера–Капелли.

Замечание 2. Если элементы какой-либо строки матрицы системы и правая часть в результате преобразований стали равными нулю, то СЛАУ совместна, но имеет бесконечное множество решений, получающихся с помощью метода Гаусса для СЛАУ порядка r, где r — ранг матрицы исходной СЛАУ.

Замечание 3. В результате прямого хода метода Гаусса можно вычислить определитель матрицы A исходной СЛАУ:

$$\det A = a_{11}a^1_{22}a^2_{33} \cdot \ldots \cdot a^{n-1}_{nn}.$$

При этом в случае перестановки строк в процессе прямого хода необходимо учитывать соответствующие перемены знаков.

Замечание 4. Метод Гаусса можно применить для обращения невырожденной ($\det A \neq 0$) матрицы.

Действительно, пусть требуется обратить невырожденную матрицу $A = [a_{ij}]$, $i, j = \overline{1, n}$. Тогда, обозначив $A^{-1} = X$, $X = [x_{ij}]$, $i, j = \overline{1, n}$, можно выписать матричное уравнение $AX = E$, где E — единичная матрица

$$E = \begin{bmatrix} 1\ 0 \ldots 0 \\ 0\ 1 \ldots 0 \\ \ldots\ldots\ldots \\ 0\ 0 \ldots 1 \end{bmatrix},$$

на основе которого можно записать цепочку СЛАУ

$$A \cdot \begin{pmatrix} x_{11} \\ x_{21} \\ \ldots \\ x_{n1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \ldots \\ 0 \end{pmatrix}, \ A \cdot \begin{pmatrix} x_{12} \\ x_{22} \\ \ldots \\ x_{n2} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \ldots \\ 0 \end{pmatrix}, \ \ldots, \ A \cdot \begin{pmatrix} x_{1n} \\ x_{2n} \\ \ldots \\ x_{nn} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 1 \end{pmatrix},$$

каждую из которых можно решить методом Гаусса. При этом, поскольку верхняя треугольная матрица для всех этих СЛАУ будет одной и той же, то метод Гаусса применяется один раз. Для этого строится следующая расширенная матрица с контрольными суммами:

$$\begin{array}{c|c|c}
x_{1n}\ x_{2n}\ \ldots\ x_{nn} & & \\
\ldots\ldots\ldots\ldots\ldots & & \\
x_{12}\ x_{22}\ \ldots\ x_{n2} & & \\
x_{11}\ x_{21}\ \ldots\ x_{n1} & b^1\ b^2\ \cdot\ b^n & k_\Sigma \\
\hline
a_{11}\ a_{12}\ \ldots\ a_{1n} & 1\ 0 \ldots 0 & \Sigma_1 \\
a_{21}\ a_{22}\ \ldots\ a_{2n} & 0\ 1 \ldots 0 & \Sigma_2 \\
\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots & \ldots \\
a_{n1}\ a_{n2}\ \ldots\ a_{nn} & 0\ 0 \ldots 1 & \Sigma_n
\end{array}$$

В результате применения $(n-1)$-го шага метода Гаусса получаем

$$\begin{array}{cccc|cccc|c}
x_{1n} & x_{2n} & \ldots & x_{nn} & & & & & \\
\multicolumn{4}{c|}{\ldots\ldots\ldots\ldots\ldots} & & & & & \\
x_{12} & x_{22} & \ldots & x_{n2} & & & & & \\
x_{11} & x_{21} & \ldots & x_{n1} & b^1 & b^2 & \cdots & b^n & k_\Sigma \\
\hline
a_{11} & a_{12} & \ldots & a_{1n} & b_{11} & b_{12} & \ldots & b_{1n} & \Sigma_1 \\
 & a^1_{22} & \ldots & a^1_{2n} & b^1_{21} & b^1_{22} & \ldots & b^1_{2n} & \Sigma^1_2 \\
\multicolumn{4}{c|}{\ldots\ldots\ldots\ldots\ldots} & \multicolumn{4}{c|}{\ldots\ldots\ldots\ldots\ldots\ldots} & \ldots \\
0 & 0 & \ldots & a^{n-1}_{nn} & b^{n-1}_{n1} & b^{n-1}_{n2} & \ldots & b^{n-1}_{nn} & \Sigma^{n-1}_n
\end{array}$$

При этом первый столбец $(\,x_{11}\ x_{21}\ \ldots\ x_{n1}\,)^T$ обратной матрицы определяется в обратном ходе метода Гаусса с правой частью b^1, столбец $(\,x_{12}\ x_{22}\ \ldots\ x_{n2}\,)^T$ — с правой частью b^2 и т. д. Столбец $(\,x_{1n}\ x_{2n}\ \ldots\ x_{nn}\,)^T$ определяется с правой частью b^n.

Пример 2.1. Методом Гаусса решить СЛАУ

$$\begin{cases} 10x_1 + x_2 + x_3 = 12, \\ 2x_1 + 10x_2 + x_3 = 13, \\ 2x_1 + 2x_2 + 10x_3 = 14. \end{cases}$$

Решение.

Прямой ход:

$$\begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & b & k_\Sigma \end{array} \\ \left(\begin{array}{ccc|c|c} \mathbf{10} & 1 & 1 & 12 & 24 \\ 2 & 10 & 1 & 13 & 26 \\ 2 & 2 & 10 & 14 & 28 \end{array}\right) \end{array} \begin{array}{l} (-2/10);\ (-2/10) \\ \xrightarrow[\text{1-й шаг}]{} \\ \end{array}$$

$$\xrightarrow[\text{1-й шаг}]{} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & b & k_\Sigma \end{array} \\ \left(\begin{array}{ccc|c|c} 10 & 1 & 1 & 12 & 24 \\ 0 & \mathbf{9{,}8} & 0{,}8 & 10{,}6 & 21{,}2 \\ 0 & 1{,}8 & 9{,}8 & 11{,}6 & 23{,}2 \end{array}\right) \end{array} (-1{,}8/9{,}8) \xrightarrow[\text{2-й шаг}]{}$$

$$\xrightarrow[\text{2-й шаг}]{} \begin{pmatrix} 10 & 1 & 1 & 12 & 24 \\ 0 & 9{,}8 & 0{,}8 & 10{,}6 & 21{,}2 \\ 0 & 0 & 9{,}653 & 9{,}653 & 19{,}31 \end{pmatrix}.$$

(столбцы: x_1, x_2, x_3 | b | k_Σ)

Обратный ход:

$$9{,}653x_3 = 9{,}653, \qquad x_3 = 1$$

$$9{,}8x_2 + 0{,}8x_3 = 10{,}6, \qquad x_2 = 1$$

$$10x_1 + x_2 + x_3 = 12, \qquad x_1 = 1.$$

Ответ: $x_1 = x_2 = x_3 = 1$.

Пример 2.2. Методом Гаусса вычислить определитель матрицы и обратить матрицу СЛАУ из примера 2.1

$$A = \begin{pmatrix} 10 & 1 & 1 \\ 2 & 10 & 1 \\ 2 & 2 & 10 \end{pmatrix}.$$

Решение.

$$\det A = 10 \cdot 9{,}8 \cdot 9{,}65 = 945{,}994 \ \ (\text{точное значение } 946).$$

Прямой ход:

x_{13} x_{23} x_{33}

x_{12} x_{22} x_{32}

x_{11} x_{21} x_{31} b^1 b^2 b^3 k_Σ

$$\begin{pmatrix} \mathbf{10} & 1 & 1 & 1 & 0 & 0 & 13 \\ 2 & 10 & 1 & 0 & 1 & 0 & 14 \\ 2 & 2 & 10 & 0 & 0 & 1 & 15 \end{pmatrix} \ (-2/10);\ (-2/10) \qquad \xrightarrow[\text{1-й шаг}]{}$$

$$\begin{matrix} x_{13} & x_{23} & x_{33} & & & & \\ x_{12} & x_{22} & x_{32} & & & & \\ x_{11} & x_{21} & x_{31} & b^1 & b^2 & b^3 & k_\Sigma \end{matrix}$$

$$\left(\begin{array}{ccc|ccc|c} 10 & 1 & 1 & 1 & 0 & 0 & 13 \\ 0 & \mathbf{9{,}8} & 0{,}8 & -0{,}2 & 1 & 0 & 11{,}4 \\ 0 & 1{,}8 & 9{,}8 & -0{,}2 & 0 & 1 & 12{,}4 \end{array}\right) \; (-1{,}8/9{,}8) \xrightarrow[\text{2-й шаг}]{}$$

$$\begin{matrix} x_{13} & x_{23} & x_{33} & & & & \\ x_{12} & x_{22} & x_{32} & & & & \\ x_{11} & x_{21} & x_{31} & b^1 & b^2 & b^3 & k_\Sigma \end{matrix}$$

$$\xrightarrow[\text{2-й шаг}]{} \left(\begin{array}{ccc|ccc|c} 10 & 1 & 1 & 1 & 0 & 0 & 13 \\ 0 & 9{,}8 & 0{,}8 & -0{,}2 & 1 & 0 & 11{,}4 \\ 0 & 0 & 9{,}653 & -0{,}163 & -0{,}184 & 1 & 10{,}31 \end{array}\right).$$

Обратный ход:

$$\begin{cases} 9{,}653x_{31} = -0{,}163, \\ 9{,}8x_{21} + 0{,}8x_{31} = -0{,}2, \\ 10x_{11} + x_{21} + x_{31} = 1; \end{cases} \qquad \begin{cases} 9{,}653x_{32} = -0{,}184, \\ 9{,}8x_{22} + 0{,}8x_{32} = 1, \\ 10x_{12} + x_{22} + x_{32} = 0; \end{cases}$$

$$\begin{cases} 9{,}653x_{33} = 1, \\ 9{,}8x_{23} + 0{,}8x_{33} = 0, \\ 10x_{13} + x_{23} + x_{33} = 0. \end{cases}$$

Отсюда $A^{-1} = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} = \begin{pmatrix} 0{,}104 & -0{,}0085 & -0{,}0095 \\ -0{,}019 & 0{,}104 & -0{,}0085 \\ -0{,}0169 & -0{,}019 & 0{,}104 \end{pmatrix}.$

Проверка:

$$A \cdot A^{-1} = \begin{pmatrix} 10 & 1 & 1 \\ 2 & 10 & 1 \\ 2 & 2 & 10 \end{pmatrix} \cdot \begin{pmatrix} 0{,}104 & -0{,}0085 & -0{,}0095 \\ -0{,}019 & 0{,}104 & -0{,}0085 \\ -0{,}0169 & -0{,}019 & 0{,}104 \end{pmatrix} =$$

$$= \begin{pmatrix} 1{,}004 & 0 & 0{,}0005 \\ 0{,}001 & 1{,}004 & 0 \\ 0{,}001 & 0{,}001 & 1{,}004 \end{pmatrix},$$

т. е. с точностью до ошибок округления получена единичная матрица.

УПРАЖНЕНИЯ.

2.1. Методом Гаусса решить СЛАУ:

а) $\begin{cases} 10x_1 - 3x_2 - 2x_3 = 5, \\ x_1 + 10x_2 - 4x_3 = 7, \\ 2x_1 - 3x_2 + 10x_3 = 9; \end{cases}$ б) $\begin{cases} 10x_1 + x_2 - 3x_3 = 8, \\ -2x_1 + 10x_2 - x_3 = 7, \\ -x_1 - 3x_2 + 10x_3 = 6. \end{cases}$

2.2. Решить примеры из задания 2.1 методом Гаусса, приведя матрицы СЛАУ к нижнему треугольному виду.

2.3. Методом Гаусса вычислить определители матриц и обратить матрицы СЛАУ из задания 2.1.

2.4. Показать с помощью метода Гаусса, что СЛАУ

$$\begin{cases} 5x_1 - x_2 + x_3 = 5, \\ x_1 + 5x_2 + x_3 = 7, \\ x_1 + 5x_2 + x_3 = 10 \end{cases}$$

несовместна.

2.5. Показать с помощью метода Гаусса, что СЛАУ

$$\begin{cases} 5x_1 - x_2 + x_3 = 5, \\ x_1 + 5x_2 + x_3 = 7, \\ 2x_1 + 10x_2 + 2x_3 = 14 \end{cases}$$

имеет бесчисленное множество решений. Найти их.

2.6. Методом Гаусса обратить матрицу

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 6 \end{pmatrix}.$$

2.1.2. Метод прогонки. Метод прогонки является одним из эффективных методов решения СЛАУ с трехдиагональными матрицами, возникающих при конечно-разностной аппроксимации задач для обыкновенных дифференциальных уравнений (ОДУ) и уравнений в частных производных второго порядка, и является частным случаем метода Гаусса. Рассмотрим следующую СЛАУ:

$$a_1 = 0 \quad \begin{cases} b_1x_1 + c_1x_2 = d_1, \\ a_2x_1 + b_2x_2 + c_2x_3 = d_2, \\ \qquad a_3x_2 + b_3x_3 + c_3x_4 = d_3, \\ \qquad \dots\dots\dots\dots\dots\dots\dots\dots \\ \qquad a_{n-1}x_{n-2} + b_{n-1}x_{n-1} + c_{n-1}x_n = d_{n-1}, \\ \qquad\qquad a_nx_{n-1} + b_nx_n = d_n, \quad c_n = 0, \end{cases} \tag{2.1}$$

решение которой будем искать в виде

$$x_i = A_ix_{i+1} + B_i, \quad i = \overline{1, n}, \tag{2.2}$$

где A_i, B_i, $i = \overline{1, n}$, — прогоночные коэффициенты, подлежащие определению. Для их определения выразим из первого уравнения СЛАУ (2.1) x_1 через x_2, получим

$$x_1 = \frac{-c_1}{b_1}x_2 + \frac{d_1}{b_1} = A_1x_2 + B_1, \tag{2.3}$$

откуда

$$A_1 = \frac{-c_1}{b_1}, \quad B_1 = \frac{d_1}{b_1}.$$

Из второго уравнения СЛАУ (2.1) с помощью (2.3) выразим x_2 через x_3, получим

$$x_2 = \frac{-c_2}{b_2 + a_2A_1}x_3 + \frac{d_2 - a_2B_1}{b_2 + a_2A_1} = A_2x_3 + B_2,$$

откуда

$$A_2 = \frac{-c_2}{b_2 + a_2 A_1}, \quad B_2 = \frac{d_2 - a_2 B_1}{b_2 + a_2 A_1}.$$

Продолжая этот процесс, получим из i-го уравнения СЛАУ (2.1):

$$x_i = \frac{-c_i}{b_i + a_i A_{i-1}} x_{i+1} + \frac{d_i - a_i B_{i-1}}{b_i + a_i A_{i-1}},$$

следовательно:

$$A_i = \frac{-c_i}{b_i + a_i A_{i-1}}, \quad B_i = \frac{d_i - a_i B_{i-1}}{b_i + a_i A_{i-1}}.$$

Из последнего уравнения СЛАУ имеем

$$x_n = \frac{-c_n}{b_n + a_n A_{n-1}} x_{n+1} + \frac{d_n - a_n B_{n-1}}{b_n + a_n A_{n-1}} = 0 \cdot x_{n+1} + B_n,$$

т. е.

$$A_n = 0 \text{ (т.к. } c_n = 0), \quad B_n = \frac{d_n - a_n B_{n-1}}{b_n + a_n A_{n-1}} = x_n.$$

Таким образом, прямой ход метода прогонки по определению прогоночных коэффициентов A_i, B_i, $i = \overline{1, n}$, завершен. В результате прогоночные коэффициенты вычисляются по следующим формулам:

$$A_i = \frac{-c_i}{b_i + a_i A_{i-1}}, \quad B_i = \frac{d_i - a_i B_{i-1}}{b_i + a_i A_{i-1}}, \quad i = \overline{2, n-1}; \tag{2.4}$$

$$A_1 = \frac{-c_1}{b_1}, \quad B_1 = \frac{d_1}{b_1}, \text{ так как } a_1 = 0, \quad i = 1; \tag{2.5}$$

$$A_n = 0, \text{ т. к. } c_n = 0, \quad B_n = \frac{d_n - a_n B_{n-1}}{b_n + a_n A_{n-1}}, \quad i = n. \tag{2.6}$$

Обратный ход метода прогонки осуществляется в соответствии с выражением (2.2):

$$\begin{cases} x_n = A_n x_{n+1} + B_n = 0 \cdot x_{n+1} + B_n = B_n, \\ x_{n-1} = A_{n-1} x_n + B_{n-1}, \\ x_{n-1} = A_{n-2} x_{n-1} + B_{n-2}, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ x_1 = A_1 x_2 + B_1. \end{cases} \tag{2.7}$$

Формулы (2.4)–(2.7) — формулы *правой прогонки.*

Аналогично, начиная с последнего уравнения СЛАУ (2.1) можно вывести формулы *левой прогонки.*

Общее число операций в методе прогонки равно $8n + 1$, т. е. пропорционально числу уравнений. Такие методы решения СЛАУ называют *экономичными.* Для сравнения метод Гаусса требует $\frac{n}{6}(2n^2 + 9n + 1)$ операций, т. е. число операций пропорционально n^3 [1].

Пример 2.3. Методом прогонки решить СЛАУ

$$\begin{cases} 8x_1 - 2x_2 = 6, \\ -x_1 + 6x_2 - 2x_3 = 3, \\ 2x_2 + 10x_3 - 4x_4 = 8, \\ -x_3 + 6x_4 = 5. \end{cases}$$

Решение.

$$A_1 = \frac{-c_1}{b_1} = \frac{2}{8} = 0,25, \quad B_1 = \frac{d_1}{b_1} = 0,75;$$

$$A_2 = \frac{-c_2}{b_2 + a_2 A_1} = \frac{2}{6 - 1 \cdot 0,25} = 0,3478,$$
$$B_2 = \frac{d_2 - a_2 B_1}{b_2 + a_2 A_1} = \frac{(3 + 1 \cdot 0,75)}{5,75} = 0,6522;$$

$$A_3 = \frac{-c_3}{b_3 + a_3 A_2} = 0,374, \quad B_3 = \frac{d_3 - a_3 B_2}{b_3 + a_3 A_2} = 0,626;$$

$$A_4 = 0 \quad (c_4 = 0), \quad B_4 = \frac{d_4 - a_4 B_3}{b_4 + a_4 A_3} = 1,0;$$

$$x_4 = A_4 x_5 + B_4 = 1,0, \quad x_3 = A_3 x_4 + B_3 = 1,0,$$

$$x_2 = A_2 x_3 + B_2 = 1,0, \quad x_1 = A_1 x_2 + B_1 = 1,0.$$

Очень часто приходится решать замкнутые СЛАУ с трехдиагональными матрицами без краевых условий ($a_1 \neq 0$, $c_n \neq 0$),

т. е. СЛАУ вида

$$\begin{cases} a_1 x_n + b_1 x_1 + c_1 x_2 = d_1, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \quad i = \overline{2, n-1}, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_n x_{n-1} + b_n x_n + c_n x_1 = d_n, \end{cases} \tag{2.8}$$

которые при $a_1 = 0$, $c_n = 0$ совпадают с системой (2.1). Такие СЛАУ решаются с помощью формул *циклической прогонки* [2]:

$$x_i = P_i x_n + Q_i, \quad i = \overline{1, n-1}; \tag{2.9}$$

$$x_n = \frac{A_{n+1} Q_1 + B_{n+1}}{1 - C_{n+1} - A_{n+1} P_1}; \tag{2.10}$$

$$P_i = A_{i+1} P_{i+1} + C_{i+1}, \quad Q_i = A_{i+1} Q_{i+1} + B_{i+1},$$

$$P_n = 1, \quad Q_n = 0, \quad i = \overline{n-1,\ 1}; \tag{2.11}$$

$$A_{i+1} = \frac{-c_i}{b_i + a_i A_i}, \ B_{i+1} = \frac{d_i - a_i B_i}{b_i + a_i A_i}, \ C_{i+1} = \frac{-a_i C_i}{b_i + a_i A_i}, \ i = \overline{1, n},$$

$$A_1 = 0, \quad B_1 = 0, \quad C_1 = 1. \tag{2.12}$$

Сначала по формулам (2.12) вычисляются прогоночные коэффициенты A_i, B_i, C_i, $i = \overline{1, n+1}$, прямого хода. Затем по формулам (2.11) вычисляются прогоночные коэффициенты обратного хода: P_i, Q_i, $i = n, n-1, \dots, 1$. Наконец, по (2.10) вычисляется x_n, а затем по формулам (2.9) — x_i, $i = \overline{1, n-1}$.

Пример 2.4. Методом циклической прогонки решить СЛАУ

$$\begin{cases} 2x_4 + 6x_1 - 2x_2 = 6, \\ -x_1 + 6x_2 - 2x_3 = 3, \\ 2x_2 + 10x_3 - 4x_4 = 8, \\ -x_3 + 6x_4 + x_1 = 6. \end{cases}$$

Решение.

1) $A_1 = 0, \quad B_1 = 0, \quad C_1 = 1;$

$$A_2 = \frac{-c_1}{b_1 + a_1 A_1} = 0{,}3333, \quad B_2 = \frac{d_1 - a_1 B_1}{b_1 + a_1 A_1} = 1{,}0,$$

$$C_2 = \frac{-a_1 C_1}{b_1 + a_1 A_1} = -0{,}3333;$$

$$A_3 = 0{,}3529, \quad B_3 = 0{,}7059, \quad C_3 = -0{,}0588;$$

$$A_4 = 0{,}3736, \quad B_4 = 0{,}6154, \quad C_4 = 0{,}01098;$$

$$A_5 = -0{,}777, \quad B_5 = 1{,}1758, \quad C_5 = 0{,}001952.$$

2) $P_4 = 1{,}0$, $Q_4 = 0$, $P_3 = A_4 P_4 + C_4 = 0{,}3846$, $Q_3 = A_4 Q_4 + B_4 = 0{,}6154$;

$$P_2 = A_3 P_3 + C_3 = 0{,}07693, \quad Q_2 = A_3 Q_3 + B_3 = 0{,}9231,$$

$$P_1 = A_2 P_2 + C_2 = -0{,}3077, \quad Q_1 = A_2 Q_2 + B_2 = 1{,}3077;$$

$$x_4 = \frac{A_5 Q_1 + B_5}{1 - C_5 - A_5 P_1} = 1{,}0, \quad x_3 = P_3 x_4 + Q_3 = 1{,}0,$$

$$x_2 = P_2 x_4 + Q_2 = 1{,}0, \quad x_1 = P_1 x_4 + Q_1 = 1{,}0.$$

Для устойчивости метода прогонки (2.4)–(2.7) достаточно выполнения следующих условий [2]:

$$|b_i| \geqslant |a_i| + |c_i|\,, \; i = \overline{1, n}, \; a_i \neq 0, \; i = \overline{2, n}, \; c_i \neq 0, \; i = \overline{1, n-1}, \tag{2.13}$$

причем строгое неравенство имеет место хотя бы при одном i. Здесь устойчивость понимается в смысле ненакопления погрешности вектора неизвестных оператором прогонки при малых погрешностях входных данных (правых частей и элементов матрицы СЛАУ). Аналогично и для циклической прогонки. При выполнении (2.13) прогоночные коэффициенты A_i (A_i, C_i, P_i в методе циклической прогонки), $i = \overline{1, n}$, не содержащие правых частей d_i СЛАУ, по модулю меньше единицы: $|A_i| < 1$ в методе прогонки и $|A_i| < 1$, $|B_i| < 1$, $|C_i| < 1$ в методе циклической прогонки.

УПРАЖНЕНИЯ.

2.7. Для СЛАУ (2.1) составить формулы левой прогонки в направлении от $i = n$ ($c_n = 0$) к $i = 1$ ($a_1 = 0$).

2.8. Для задачи $a_i x_{i-1} - b_i x_i + c_i x_{i+1} = d_i$, $|b_i| \geqslant |a_i| + |c_i|$, $a_i \neq 0$, $c_i \neq 0$, $i = \overline{1, n-1}$; $x_0 = c_0 x_1 + d_0$, $i = 0$; $x_n = c_n x_{n-1} + d_n$, $i = n$, составить формулы правой и левой прогонки. Показать, что прогоночные коэффициенты, не содержащие правых частей, меньше единицы.

2.9. Для СЛАУ $\frac{1}{h^2} u_{i-1} - \frac{2}{h^2} u_i + \frac{1}{h^2}(u_{i+1} + u_i) = h^2$, $i = \overline{1, n}$, $u_0 = u_{n+1} = 1$, $h = 0,1$, $n = 10$ достаточные условия устойчивости (2.13) не выполняются. Является ли метод прогонки устойчивым?

2.10. Используя метод левой прогонки, выписать формулы точного решения системы $a_i x_{i-1} - b_i x_i + c_i x_{i+1} = d_i$, $|b_i| \geqslant |a_i| + |c_i|$, $a_i \neq 0$, $c_i \neq 0$, $i = \overline{1, n-1}$; $x_0 = c_0 x_1 + \sigma x_0^4 + d_0$, $i = 0$; $x_n = c_n x_{n-1} + d_n$, $i = n$, где нулевое уравнение является нелинейным; $\sigma = 10^{-10}$.

2.11. Доказать, что метод прогонки для СЛАУ (2.1) есть метод Гаусса.

2.12. Записать формулы циклической прогонки для СЛАУ $a_i u_{i-1} - b_i u_i + c_i u_{i+1} = d_i$, $i = \overline{0, n}$; $u_{-1} = u_n$, $u_{n+1} = u_0$; $|b_i| \geqslant |a_i| + |c_i|$, $a_i \neq 0$, $c_i \neq 0$ (строгое неравенство имеет место хотя бы для одного i).

2.13. Найти решение разностной системы $u_{i-1} - 2u_i + u_{i+1} = 0$, $i = \overline{1, n-1}$; $u_0 = 1$; $u_n = 2$; $n = 6$.

2.14. Найти решение разностной системы $-u_{i-1} + 2u_i - u_{i+1} = h^2 \sin(ih)$, $i = \overline{1, n-1}$; $u_0 = u_1 - h$; $u_n = 1$; $nh = \pi/2$. Определить погрешность путем сравнения с решением задачи $\frac{d^2 u}{dx^2} + \sin x = 0$; $\frac{du(0)}{dx} = 1$; $u\left(\frac{\pi}{2}\right) = 1$.

2.1.3. Обоснование метода прогонки. Метод прогонки содержит операцию деления и следовательно возможно накопление ошибок при увеличении числа уравнений в СЛАУ (2.1) или числа узлов сетки при конечно-разностной аппроксимации краевых задач.

Поэтому необходимо гарантировать *корректность*, т. е. выполнение условия

$$b_i + a_i A_{i-1} \neq 0, \quad i = \overline{1, n}, \tag{2.14}$$

и *устойчивость*, т. е, ненакопление ошибок при увеличении числа уравнений СЛАУ (2.1).

Пусть прогоночные коэффициенты A_i, B_i, $i = \overline{1, n}$, вычислены точно, а при вычислении x_n допущена ошибка ε_n: $\widehat{x}_n = x_n + \varepsilon_n$. Тогда из (2.7) имеем следующие равенства:

$$\widehat{x}_i = A_i \widehat{x}_{i+1} + B_i, \quad i = n-1, \ldots, 1,$$

$$x_i = A_i x_{i+1} + B_i, \quad i = n-1, \ldots, 1.$$

Вычитая из первого равенства второе, получаем

$$\varepsilon_i = A_i \varepsilon_{i+1}, \quad i = n-1, n-2, \ldots, 1,$$

т. е., если выполняются условия

$$|A_i| < 1, \quad i = \overline{1, n}, \tag{2.15}$$

то алгоритм метода прогонки не накапливает ошибок и является *устойчивым*.

Для корректности и устойчивости метода прогонки, т. е. для реализации неравенств (2.14), (2.15), существует следующая лемма.

Лемма 2.1 (*достаточное условие корректности и устойчивости метода прогонки*):

Пусть коэффициенты СЛАУ (2.1) удовлетворяют условиям

$$|a_i| \geqslant 0, \quad |b_i| > 0, \quad |c_i| \geqslant 0, \quad i = \overline{1, n},$$

$$|b_i| \geqslant |a_i| + |c_i|, \quad i = \overline{2, n-1}; \tag{2.16}$$

$$|b_1| \geqslant |c_1|, \quad |b_n| \geqslant |a_n|, \tag{2.17}$$

причем хотя бы в одном из неравенств (2.16) или (2.17) выполняется строгое неравенство, т. е. матрица СЛАУ (2.1) имеет диагональное преобладание. Тогда имеют место неравенства (2.14) и (2.15), гарантирующие корректность и устойчивость метода прогонки.

Действительно, из (2.17) имеем неравенство $|A_1| = \dfrac{|c_1|}{|b_1|} \leqslant 1$, а из (2.16) — неравенства

$$|A_i| = \frac{|c_i|}{|b_i + a_i A_{i-1}|} \leqslant \frac{|b_i| - |a_i|}{|b_i + a_i A_{i-1}|} \leqslant \frac{|b_i| - |a_i|\,|A_{i-1}|}{|b_i + a_i A_{i-1}|} \leqslant 1,$$

$$i = \overline{2,\, n-1}.$$

Кроме этого, из условия (2.16) леммы имеем

$$|b_i + a_i A_{i-1}| \geqslant |b_i| - |a_i|\,|A_{i-1}| \geqslant |a_i| + |c_i| - |a_i|\,|A_{i-1}| \geqslant |c_i| > 0,$$

т. е. $b_i + a_i A_{i-1} \neq 0$, что и требовалось доказать.

2.1.4. Матричная прогонка. При численном решении многомерных задач математической физики, например двумерных задач для уравнений Лапласа, теплопроводности и т. п., результирующая СЛАУ имеет пятидиагональную матрицу, которую можно представить следующей системой векторно-матричных уравнений:

$$\begin{gathered} A_i X_{i-1} + B_i X_i + C_i X_{i+1} = F_i, \quad i = \overline{0, N}, \\ A_0 = C_N = \Theta, \end{gathered} \tag{2.18}$$

где Θ — нулевая матрица.

Здесь A_i, B_i, C_i, $i = \overline{0, N}$, — квадратные матрицы размерности $M \times M$; X_i, F_i, $i = \overline{0, N}$, — соответственно искомые векторы и векторы правых частей размера M. При этом должно выполняться условие

$$\det B_i \neq 0, \quad i = \overline{0, N}.$$

Так же как и в методе скалярной прогонки, решение ищется в виде

$$X_i = P_i X_{i+1} + Q_i, \quad i = \overline{0, N}, \tag{2.19}$$

где P_i, Q_i, $i = \overline{0, N}$, — соответственно прогоночные матрицы и прогоночные векторы, подлежащие определению. Для их определения выразим из нулевого векторно-матричного уравнения системы (2.18) X_0 через X_1, получим

$$X_0 = -B_0^{-1} C_0 X_1 + B_0^{-1} F_0 = P_0 X_1 + Q_0, \tag{2.20}$$

откуда

$$P_0 = -B_0^{-1} C_0, \quad Q_0 = B_0^{-1} F_0.$$

Далее из первого векторно-матричного уравнения системы (2.18) выразим X_1 через X_2 с использованием (2.20), получим

$$X_1 = -(B_1 + A_1 P_0)^{-1} C_1 X_2 + (B_1 + A_1 P_0)^{-1} (F_1 - A_1 Q_0) = \\ = P_1 X_2 + Q_1,$$

откуда

$$P_1 = -(B_1 + A_1 P_0)^{-1} C_1,$$

$$Q_1 = (B_1 + A_1 P_0)^{-1}(F_1 - A_1 Q_0).$$

Продолжая этот процесс, получим прогоночные матрицы P_i и прогоночные векторы Q_i в следующем виде:

$$\begin{gathered} P_i = -(B_i + A_i P_{i-1})^{-1} C_i, \\ Q_i = (B_i + A_i P_{i-1})^{-1}(F_i - A_i Q_{i-1}), \quad i = \overline{0, N}, \\ A_0 = \Theta, \quad C_N = \Theta. \end{gathered} \tag{2.21}$$

Прямой ход метода матричной прогонки завершен. Искомые векторы X_i определяются из соотношений (2.19):

$$\begin{gathered} X_N = P_N X_{N+1} + Q_N = Q_N \ \ (P_N = \Theta), \\ X_{N-1} = P_{N-1} X_N + Q_{N-1}, \\ \dots\dots\dots\dots\dots\dots \\ X_0 = P_0 X_1 + Q_0, \end{gathered} \tag{2.22}$$

называемых обратным ходом метода матричной прогонки.

Матричная прогонка неэкономична, так как она включает в себя операцию обращения матриц.

Алгоритм (2.21), (2.22) *устойчив и корректен*, если выполнены следующие условия:

$$\begin{gathered} \det B_i \neq 0, \quad i = \overline{0, N}; \\ A_i \neq \Theta, \quad i = \overline{1, N}; \\ C_i \neq \Theta, \quad i = \overline{0, N-1}; \\ \|B_0^{-1} C_0\| < 1; \quad \|B_N^{-1} C_N\| < 1; \\ \|B_i^{-1} C_i\| + \|B_i^{-1} A_i\| < 1, \quad i = \overline{1, N-1}. \end{gathered}$$

Здесь норма матрицы может быть выбрана любой (см. следующий раздел).

2.1.5. Нормы векторов и матриц. Для исследования сходимости численных методов решения задач линейной алгебры вводятся понятия нормы векторов и матриц.

Нормой вектора $x = (x_1, x_2, \ldots, x_n)^T$ (обозначают $\|x\|$) в n-мерном вещественном пространстве векторов $x \in R^n$ называют неотрицательное число, вычисляемое с помощью компонент вектора и обладающее следующими свойствами:

а) $\|x\| \geqslant 0$ ($\|x\| = 0$ тогда и только тогда, когда x — нулевой вектор $x = \vartheta$);

б) $\|\alpha \cdot x\| = |\alpha| \cdot \|x\|$ для любых чисел α (действительных или комплексных);

в) $\|x + y\| \leqslant \|x\| + \|y\|$.

Нормой матрицы $A_{n\times n}$ (обозначается $\|A\|$) с вещественными элементами в пространстве матриц называют неотрицательное число, вычисляемое с помощью элементов матрицы и обладающее следующими свойствами:

а) $\|A\| > 0$ ($\|A\| = 0$ тогда и только тогда, когда А — нулевая матрица $A = \Theta$);

б) $\|\alpha \cdot A\| = |\alpha| \cdot \|A\|$ для любых действительных и комплексных чисел α;

в) $\|A + B\| \leqslant \|A\| + \|B\|$;

г) $\|A \cdot B\| \leqslant \|A\| \cdot \|B\|$ для всех $n \times n$-матриц A и B рассматриваемого пространства.

Как видно из определения норм векторов и матриц (определения аналогичны, за исключением последнего свойства нормы матрицы), норма матриц должна быть согласована с нормой векторов. Это согласование осуществляется связью

$$\|Ax\| \leqslant \|A\|\,\|x\|\,. \tag{2.23}$$

Наиболее употребительными являются следующие нормы векторов:

$$\|x\|_1 = \max_i |x_i|\,, \tag{2.24}$$

$$\|x\|_2 = \sum_{i=1}^{n} |x_i|, \tag{2.25}$$

$$\|x\|_3 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{(x, x)}\,. \tag{2.26}$$

Согласованными с ними с помощью связи (2.23) нормами матриц будут соответственно:

$$\|A\|_1 = \max_i \sum_{j=1}^{n} |a_{ij}|, \tag{2.27}$$

$$\|A\|_2 = \max_j \sum_{i=1}^{n} |a_{ij}|, \tag{2.28}$$

$$\|A\|_3 = \sqrt{\max_i |\lambda_i|_{A^T A}}\,. \tag{2.29}$$

Можно показать [3], что нормы матриц (2.27)–(2.29) согласованы с помощью связи (2.23) с соответствующими нормами векторов (2.24)–(2.26). Под знаком квадратного корня в норме матрицы $||A||_3$ находится спектральный радиус симметрической матрицы $A^T A$, для которой, как известно, все собственные значения являются действительными.

Из линейной алгебры известно, что собственные значения матриц не превышают их норм. Действительно, из равенства $Ax = \lambda x$ и свойств норм векторов и матриц следует: $\|\lambda x\| = |\lambda|\,\|x\| = \|Ax\| \leqslant \|A\|\,\|x\|$, $|\lambda| \leqslant \|A\|$, или $\rho(A) \leqslant \|A\|$, где $\rho(A) = \max\limits_i |\lambda_i|$ — максимальное по модулю собственное значение или спектральный радиус матрицы A. Таким образом, за норму матрицы можно принять ее спектральный радиус.

Пример 2.5. Показать согласованность нормы вектора $\|x\|_1$ (2.24) и нормы матрицы $\|A\|_1$ (2.27).

Решение.

$$\|Ax\|_1 = \max_i |Ax| = \max_i \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant$$
$$\leqslant \max_i \sum_{j=1}^{n} |a_{ij} x_j| \leqslant \max_i |x_i| \max_i \sum_{j=1}^{n} |a_{ij}| =$$
$$= \|x\|_1 \max_i \sum_{j=1}^{n} |a_{ij}| \Rightarrow \frac{\|Ax\|_1}{\|x\|_1} \leqslant \max_i \sum_{j=1}^{n} |a_{ij}|\,.$$

Таким образом, условие согласования (2.23) выполнено.

Для исследования погрешностей, возникающих при решении СЛАУ, вводят понятие *числа обусловленности матрицы* $\operatorname{cond}(A)$ [4]:

$$\operatorname{cond}(A) = \|A\| \cdot \left\|A^{-1}\right\|.$$

Если в качестве нормы матрицы принять ее спектральный радиус $\max\limits_i |\lambda_i|$, то

$$\operatorname{cond}(A) = \max_i |\lambda_i| \frac{1}{\min\limits_i |\lambda_i|} \geqslant 1,$$

поскольку спектральный радиус обратной матрицы A^{-1} равен обратной величине минимального собственного значения исходной матрицы.

Число обусловленности характеризует степень зависимости относительной погрешности решения СЛАУ от погрешности входных данных (правых частей и элементов матрицы). Можно показать, что справедливы следующие неравенства:

$$\frac{\|\Delta x\|}{\|x\|} \leqslant \operatorname{cond} A \frac{\|\Delta b\|}{\|b\|}, \qquad \frac{\|\Delta x\|}{\|x\|} \leqslant \operatorname{cond} A \frac{\|\Delta A\|}{\|A + \Delta A\|}.$$

Таким образом, чем больше число обусловленности, тем сильнее влияние погрешности входных данных на конечный результат. Матрица считается плохо обусловленной, если $\operatorname{cond}(A) \gg$ $\gg 1$. Чем ближе $\operatorname{cond}(A)$ к 1, тем матрица лучше обусловлена. Примером может служить ортогональная матрица.

Пример 2.6. Вычислить число обусловленности для матрицы

$$A = \begin{bmatrix} 1{,}0 & 0{,}99 \\ 0{,}99 & 0{,}98 \end{bmatrix}.$$

Решение.

Для этой матрицы $\det A = 10^{-4} \neq 0$; $A^{-1} = 10^4 \times$ $\times \begin{bmatrix} 0,98 & -0,99 \\ -0,99 & 1,0 \end{bmatrix}$; $\|A\|_1 = 1,99$; $\left\|A^{-1}\right\|_1 = 1,99 \cdot 10^4$; $\operatorname{cond}(A) =$ $= 39\,601$.

УПРАЖНЕНИЯ.

2.15. Используя неравенство Коши–Буняковского $\left|x^T \mathrm{y}\right| \leqslant$ $\leqslant \|x\| \, \|y\|$, доказать второе и третье свойства для норм вектора, определяемых соотношениями (2.24)–(2.26).

2.16. Доказать согласованность норм (2.25) вектора и (2.28) матрицы [3].

2.17. Согласованы ли нормы (2.26) и (2.29)?

2.18. Доказать, что для любой $(n \times n)$-матрицы A $\max\limits_{i,j} |a_{ij}| \leqslant \|A\| \leqslant n \cdot \max\limits_{ij} |a_{ij}|$.

2.19. Доказать, что если матрица ортогональна ($A^{-1} = A^T$), то $\text{cond}(A) = 1$.

2.20. Пусть

$$A = \begin{pmatrix} 6 & 13 & -17 \\ 13 & 29 & -38 \\ -17 & -38 & 50 \end{pmatrix}; \quad A^{-1} = \begin{pmatrix} 6 & -4 & -1 \\ -4 & 11 & 7 \\ -1 & 7 & 5 \end{pmatrix}; \quad \begin{matrix} \lambda_1 \approx 0{,}0588, \\ \lambda_2 \approx 0{,}2007, \\ \lambda_3 \approx 84{,}74. \end{matrix}$$

Найти $\|A\|$, $\|A^{-1}\|$, $\text{cond}(A)$.

2.1.6. Итерационные методы решения СЛАУ. Метод простых итераций. При большом числе уравнений ($\sim$ 100 и более) прямые методы решения СЛАУ (за исключением метода прогонки) становятся труднореализуемыми на ЭВМ, прежде всего из-за сложности хранения и обработки матриц большой размерности.

Методы последовательных приближений, в которых при вычислении последующего приближения решения используются предыдущие, уже известные приближенные решения, называются *итерационными*.

В итерационных методах решение может быть вычислено за бесконечное число итераций (приближений), а поскольку это невозможно, то, останавливая процесс вычислений на какой-либо итерации, необходимо уметь оценивать погрешность метода итераций.

Рассмотрим СЛАУ

$$\begin{cases} a_{11}\underline{x_1} + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}\underline{x_2} + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}\underline{x_n} = b_n \end{cases} \tag{2.30}$$

с невырожденной матрицей ($\det A \neq 0$).

Приведем СЛАУ к эквивалентному виду:

$$\begin{cases} x_1 = \beta_1 + \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1n}x_n, \\ x_2 = \beta_2 + \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = \beta_n + \alpha_{n1}x_1 + \alpha_{n2}x_2 + \ldots + \alpha_{nn}x_n, \end{cases} \tag{2.31}$$

или, в векторно-матричной форме:

$$x = \beta + \alpha x, \tag{2.32}$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix}.$$

Такое приведение может быть выполнено различными способами. Одним из наиболее распространенных является следующий.

Разрешим систему (2.30) относительно подчёркнутых неизвестных при ненулевых диагональных элементах $a_{ii} \neq 0$, $i = \overline{1, n}$ (если какой-либо коэффициент на главной диагонали равен нулю, достаточно соответствующее уравнение поменять местами с любым другим уравнением). Получим следующие выражения для компонентов вектора β и матрицы α эквивалентной системы:

$$\beta_i = \frac{b_i}{a_{ii}}; \quad \alpha_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}}, \ i, j = \overline{1, n}, \ i \neq j; \\ \quad 0, \quad i = j, \ i = \overline{1, n}. \end{cases}$$

Отметим, что существуют и другие способы приведения СЛАУ к эквивалентному виду.

В качестве нулевого приближения $x^{(0)}$ вектора неизвестных примем вектор правых частей $x^{(0)} = \beta$, или $\left(x_1^{(0)} \ x_2^{(0)} \ \ldots \ x_n^{(0)} \right)^T = (\,\beta_1 \ \beta_2 \ \ldots \ \beta_n\,)^T$. Тогда итерационную последовательность векторов

$$\begin{cases} x^{(0)} = \beta, \\ x^{(1)} = \beta + \alpha x^{(0)}, \\ x^{(2)} = \beta + \alpha x^{(1)}, \\ \dots\dots\dots\dots\dots\dots \\ x^{(k)} = \beta + \alpha x^{(k-1)}. \end{cases} \tag{2.33}$$

называют *методом простых итераций.*

При решении СЛАУ (2.30) методом простых итераций необходимо отвечать на следующие вопросы.

1. Сходится ли последовательность $x^{(0)}, x^{(1)}, x^{(2)}, \dots$ $\dots, x^{(k)}, \dots$, т.е. существует ли $\lim\limits_{k\to\infty} x^{(k)}$? Если предел существует $(\lim\limits_{k\to\infty} x^{(k)} = \xi)$, то необходимо ответить на второй вопрос.

2. Является ли предельный вектор ξ решением СЛАУ (2.31) или (2.32)? Если ξ — решение, т.е. (2.32) удовлетворяется тождественно $\xi = \beta + \alpha\xi$, то необходимо ответить на третий вопрос.

3. Так как бесконечное число итераций осуществить невозможно, то, останавливая процесс на k-ой итерации, необходимо определить: какова погрешность по норме $\left\|\xi - x^k\right\|$? Или из условия заданной точности ε, определить, каково число итераций k, удовлетворяющее этой точности?

Для ответа на первые два вопроса существует

Теорема 2.1. (*достаточное условие сходимости метода простых итераций*) [1].

Метод простых итераций (2.33) *сходится к единственному решению СЛАУ* (2.32) (*а следовательно, и к решению исходной СЛАУ* (2.30)) *при любом начальном приближении* x^0, *если какая-либо норма матрицы* α *эквивалентной системы меньше единицы:*

$$\|\alpha\| < 1 \ \ \forall \|\alpha\| \ \text{ и } \forall x^0.$$

Доказательство.

Запишем последовательность (2.33) в виде

$$x^{(0)} = \beta,$$

$$x^{(1)} = \beta + \alpha\beta = (E + \alpha)\beta,$$

$$x^{(2)} = \beta + \alpha(E + \alpha)\beta = (E + \alpha + \alpha^2)\beta,$$

$$x^{(3)} = \beta + \alpha(E + \alpha + \alpha^2)\beta = (E + \alpha + \alpha^2 + \alpha^3)\beta,$$
$$\dots\dots\dots\dots \qquad (2.34)$$
$$x^{(k)} = (E + \alpha + \alpha^2 + \ldots + \alpha^k)\beta,$$

где E — единичная матрица с нормой $\|E\| = 1$.

Переходя вначале в равенствах (2.34) к норме с использованием свойств норм векторов и матриц:

$$\left\|x^{(k)}\right\| = \left\|(E + \alpha + \alpha^2 + \ldots + \alpha^k)\beta\right\| \leqslant \left\|E + \alpha + \alpha^2 + \ldots \right.$$
$$\left. \ldots + \alpha^k\right\| \cdot \|\beta\| \leqslant \left(\|E\| + \|\alpha\| + \|\alpha\|^2 + \ldots + \|\alpha\|^k\right) \cdot \|\beta\|,$$

а затем к пределу при $k \to \infty$, получим

$$\lim_{k\to\infty} \left\|x^{(k)}\right\| \leqslant \left[\lim_{k\to\infty} \left(1 + \|\alpha\| + \ldots + \|\alpha\|^k\right)\right] \cdot \|\beta\|.$$

В квадратных скобках этого неравенства стоит ряд геометрической прогрессии со знаменателем $\|\alpha\|$, который легко суммируется:

$$\lim_{k\to\infty} \left\|x^{(k)}\right\| \leqslant \lim_{k\to\infty} \frac{1 - \|\alpha\|^k}{1 - \|\alpha\|} \|\beta\|.$$

При $\|\alpha\| \geqslant 1$ данный предел не существует, а при $\|\alpha\| < 1$ сумма бесконечно убывающей геометрической прогрессии равна $\dfrac{1}{1 - \|\alpha\|}$. Таким образом,

$$\lim_{k\to\infty} \left\|x^{(k)}\right\| \leqslant \frac{\|\beta\|}{1 - \|\alpha\|},$$

т. е. предел итерационной последовательности (2.33) существует. Обозначим этот предел через ξ: $\lim\limits_{k\to\infty} x^{(k)} = \xi$.

Ответим на 2-й вопрос: является ли этот предельный вектор ξ решением эквивалентной СЛАУ (2.32), а следовательно, и исходной СЛАУ (2.30)? Для этого в равенствах (2.34) перейдем к пределу при $k \to \infty$:

$$\lim_{k\to\infty} x^{(k)} = \left[\lim_{k\to\infty} \left(E + \alpha + \alpha^2 + \ldots + \alpha^k\right)\right] \cdot \beta. \qquad (2.35)$$

В квадратных скобках равенства (2.35) стоит матричный ряд, формально являющийся биномиальным рядом с суммой: $S = (E-\alpha)^{-1}$.

Следовательно,

$$\lim_{k\to\infty} x^{(k)} = \xi = (E-\alpha)^{-1}\cdot\beta,$$

откуда получаем следующую цепочку равенств:

$$(E-\alpha)\,\xi = \beta \Rightarrow \xi - \alpha\xi = \beta, \quad \text{или} \quad \xi = \beta + \alpha\xi. \tag{2.36}$$

Сравнивая (2.36) и (2.32), заключаем, что ξ — решение СЛАУ (2.32), а следовательно, и СЛАУ (2.30).

Ответим на 3-й вопрос. Покажем, что если процесс итераций остановить на k-й итерации, то норма погрешности $\left\|\xi - x^{(k)}\right\|$ будет оцениваться неравенством

$$\left\|\xi - x^{(k)}\right\| \leqslant \frac{\|\alpha\|^{k+1}}{1-\|\alpha\|}\,\|\beta\|\,. \tag{2.37}$$

Действительно, рассмотрим равенства (2.33) на двух соседних итерациях, получим

$$\begin{cases} x^{(k+1)} = \alpha x^{(k)} + \beta, \\ x^{(k)} = \alpha x^{(k-1)} + \beta, \end{cases}$$

откуда

$$\left\|x^{(k+1)} - x^{(k)}\right\| \leqslant \|\alpha\| \left\|x^{(k)} - x^{(k-1)}\right\|.$$

На основе последнего неравенства можно записать следующую цепочку неравенств:

$$\left\|x^{(k+2)} - x^{(k+1)}\right\| \leqslant \|\alpha\| \left\|x^{(k+1)} - x^{(k)}\right\|,$$

$$\left\|x^{(k+3)} - x^{(k+2)}\right\| \leqslant \|\alpha\|^2 \left\|x^{(k+1)} - x^{(k)}\right\|,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\left\|x^{(k+p)} - x^{(k+p-1)}\right\| \leqslant \|\alpha\|^{p-1} \left\|x^{(k+1)} - x^{(k)}\right\|,$$

используя которую в тождестве

$$x^{(k+p)} - x^{(k)} = \left(x^{(k+1)} - x^{(k)}\right) + \left(x^{(k+2)} - x^{(k+1)}\right) + \ldots$$
$$\ldots + \left(x^{(k+p)} - x^{(k+p-1)}\right),$$

получим неравенство

$$\left\|x^{(k+p)} - x^{(k)}\right\| \leqslant \left(1 + \|\alpha\| + \|\alpha\|^2 + \ldots + \|\alpha\|^{p-1}\right)\left\|x^{(k+1)} - x^{(k)}\right\|,$$

откуда при $p \to \infty$ приходим к неравенству (2.37) (при $\|\alpha\| < 1$):

$$\left\|\xi - x^{(k)}\right\| \leqslant \frac{1}{1 - \|\alpha\|}\left\|x^{(k+1)} - x^{(k)}\right\| \leqslant \leqslant \frac{\|\alpha\|^k}{1 - \|\alpha\|}\left\|x^{(1)} - x^{(0)}\right\| \leqslant \frac{\|\alpha\|^{k+1}}{1 - \|\alpha\|}\|\beta\|, \quad (2.38)$$

так как $x^{(1)} - x^{(0)} = \left(\beta + \alpha x^{(0)}\right) - \beta = \alpha \cdot \beta$, $x^{(0)} = \beta$.

Если задана точность ε, то

$$\frac{\|\alpha\|^{k+1}}{1 - \|\alpha\|}\|\beta\| \leqslant \varepsilon,$$

откуда получаем априорную нижнюю оценку числа итераций k при $\|\alpha\| < 1$:

$$k + 1 \geqslant \frac{\lg\varepsilon - \lg\|\beta\| + \lg(1 - \|\alpha\|)}{\lg\|\alpha\|}.$$

Это неравенство обычно дает завышенное число итераций k.

Как правило, при выполнении достаточного условия сходимости итерационный процесс останавливается при удовлетворении условию $\varepsilon^{(k)} = \dfrac{\|\alpha\|}{1 - \|\alpha\|}\left\|x^{(k)} - x^{(k-1)}\right\| \leqslant \varepsilon$ (иногда используется более простое условие $\left\|x^{(k)} - x^{(k-1)}\right\| \leqslant \varepsilon$), которое проверяется начиная с $k = 2$. После выполнения этого условия принимается $\xi \approx x^{(k)}$.

Если используется метод Якоби (выражения (2.32) для эквивалентной СЛАУ), то достаточным условием сходимости является *диагональное преобладание матрицы* A, т. е. $|a_{ii}| > \sum\limits_{j=1, i\neq j}^{n} |a_{ij}|$ $\forall i$ (для каждой строки матрицы A модули элементов, стоящих на главной диагонали, больше суммы модулей недиагональных элементов). Очевидно, что в этом случае $\|\alpha\|_1$ меньше единицы и, следовательно, итерационный процесс (2.33) сходится.

Теорема 2.2 (*необходимое и достаточное условие сходимости метода простых итераций*). *Для сходимости итерационного процесса* (2.33) *необходимо и достаточно, чтобы спектр матрицы* α *эквивалентной системы лежал внутри круга с радиусом, равным единице.*

Спектр — все собственные значения матрицы на комплексной плоскости (γ, β) (рис.2.1). Докажем только достаточность: *если спектр матрицы* α *меньше единицы, то метод простых итераций* (2.33) *сходится к единственному решению* системы (2.32).

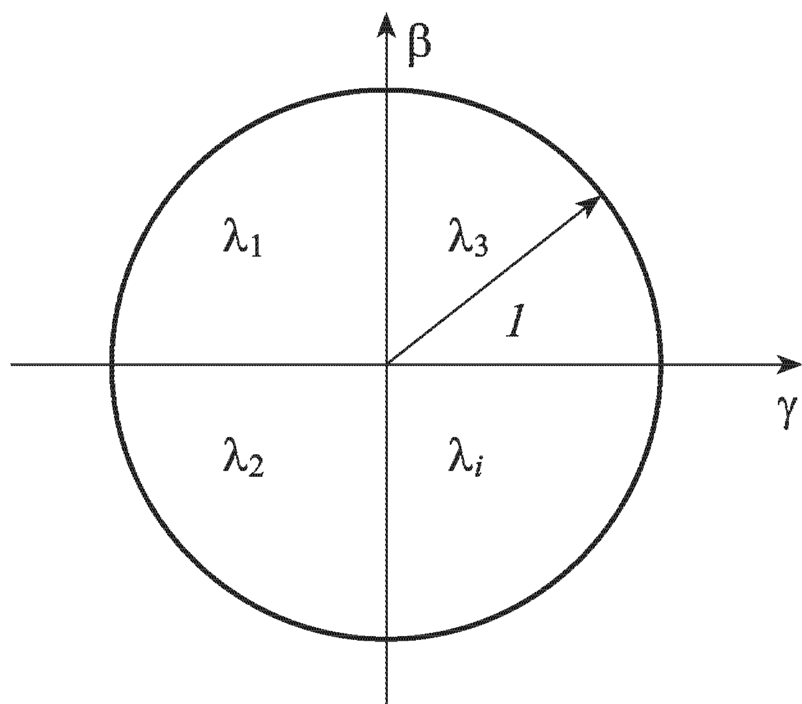


Рис. 2.1. Спектр матрицы эквивалентной системы

Доказательство.

Рассмотрим задачу на собственные значения и собственные векторы матрицы α: $\alpha y = \lambda y$.

Применяя к обеим частям этого равенства операцию нормы и учитывая, что $\max\limits_i |\lambda_i| < 1$, получим

$$\|\alpha y\| = |\lambda| \cdot |y| < \|y\|, \quad \frac{\|\alpha y\|}{\|y\|} < 1,$$

откуда, в соответствии с равенством (2.23), заключаем:

$$\|\alpha\| = \frac{\|\alpha y\|}{\|y\|} < 1,$$

что достаточно для сходимости итерационного процесса (2.33) (см. теорему 2.1).

Пример 2.7. Методом простых итераций с точностью $\varepsilon = 0,01$ решить СЛАУ

$$\begin{cases} 10x_1 + x_2 + x_3 = 12, \\ 2x_1 + 10x_2 + x_3 = 13, \\ 2x_1 + 2x_2 + 10x_3 = 14. \end{cases}$$

Решение.

Приведем СЛАУ к эквивалентному виду

$$\begin{cases} x_1 = 1{,}2 - 0{,}1x_2 - 0{,}1x_3, \\ x_2 = 1{,}3 - 0{,}2x_1 - 0{,}1x_3, \\ x_3 = 1{,}4 - 0{,}2x_1 - 0{,}2x_2 \end{cases}$$

или $x = \beta + \alpha x$, где $\alpha = \begin{pmatrix} 0 & -0{,}1 & -0{,}1 \\ -0{,}2 & 0 & -0{,}1 \\ -0{,}2 & -0{,}2 & 0 \end{pmatrix}$; $\beta =$ $= (\,1{,}2 \;\; 1{,}3 \;\; 1{,}4\,)^T$; $\|\alpha\|_1 = 0{,}4 < 1$, следовательно, достаточное условие сходимости метода простых итераций выполнено.

Итерационный процесс выглядит следующим образом:

$$x^{(0)} = \beta; \quad x^{(1)} = \beta + \alpha\beta = (\; 0{,}93 \;\; 0{,}92 \;\; 0{,}9 \;)^T;$$

$$\varepsilon^{(1)} = 0,333 > \varepsilon;$$

$$x^{(2)} = \beta + \alpha x^{(1)} = (\; 1{,}018 \;\; 1{,}024 \;\; 1{,}03 \;)^T;$$

$$\varepsilon^{(2)} = 0{,}0867 > \varepsilon$$

$$x^{(3)} = \beta + \alpha x^{(2)} = (\; 0{,}9946 \;\; 0{,}9934 \;\; 0{,}9916 \;)^T;$$

$$\varepsilon^{(3)} = 0{,}0256 > \varepsilon$$

$$x^{(4)} = \beta + \alpha x^{(3)} = (\; 1{,}0015 \;\; 1{,}00192 \;\; 1{,}0024 \;)^T;$$

$$\varepsilon^{(4)} = 0{,}0072 < \varepsilon.$$

Таким образом, вычислительный процесс завершен за 4 итерации. Отметим, что точное решение исходной СЛАУ в данном случае известно: $x^* = (\; 1 \;\; 1 \;\; 1 \;)^T$. Отсюда следует, что заданной точности $\varepsilon = 0{,}01$ удовлетворяло решение, полученное уже на третьей итерации. Но в силу использования оценочного выражения для погрешности процесс останавливается только на четвертой итерации.

Отметим также, что априорная оценка необходимого количества итераций в данной задаче дает: $k + 1 \geqslant (-2 + \lg 0{,}6 - - \lg 1{,}4)/\lg 0{,}4 = 5{,}95$, т. е. для достижения точности $\varepsilon = 0{,}01$, согласно априорной оценке, необходимо сделать не менее пяти итераций, что иллюстрирует характерную для априорной оценки тенденцию к завышению числа итераций.

2.1.7. Метод Зейделя решения СЛАУ. Метод простых итераций довольно медленно сходится. Для его ускорения существует *метод Зейделя*, заключающийся в том, что при вычислении компонента $x_i^{(k+1)}$ вектора неизвестных на $(k+1)$-й итера-

ции используются $x_1^{(k+1)}$, $x_2^{(k+1)}$, ..., $x_{i-1}^{(k+1)}$, уже вычисленные на $(k+1)$-й итерации. Значения остальных компонентов $x_i^{(k)}$, $x_{i+1}^{(k)}, \ldots, x_n^{(k)}$ берутся из предыдущей итерации. Так же как и в методе простых итераций, строится эквивалентная СЛАУ (2.31) и за начальное приближение принимается вектор правых частей: $x^{(0)} = (\ \beta_1\, \beta_2 \ldots \beta_n\ )^T$. Тогда метод Зейделя для известного вектора $\left(\ x_1^{(k)}\ x_2^{(k)} \ldots x_n^{(k)}\ \right)^T$ на k-ой итерации будет иметь вид

$$\begin{cases} x_1^{(k+1)} = \beta_1 + \alpha_{11}x_1^{(k)} + \alpha_{12}x_2^{(k)} + \ldots + \alpha_{1n}x_n^{(k)}, \\ x_2^{(k+1)} = \beta_2 + \alpha_{21}x_1^{(k+1)} + \alpha_{22}x_2^{(k)} + \ldots + \alpha_{2n}x_n^{(k)}, \\ x_3^{(k+1)} = \beta_3 + \alpha_{31}x_1^{(k+1)} + \alpha_{32}x_2^{(k+1)} + \alpha_{33}x_3^{(k)} + \ldots + \alpha_{3n}x_n^{(k)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = \beta_n + \alpha_{n1}x_1^{(k+1)} + \alpha_{n2}x_2^{(k+1)} + \ldots \\ \qquad\qquad \ldots + \alpha_{nn-1}x_{n-1}^{(k+1)} + \alpha_{nn}x_n^{(k)}. \end{cases}$$

Из этой системы видно, что $x^{(k+1)} = \beta + Bx^{(k+1)} + Cx^{(k)}$, где B — нижняя треугольная матрица с диагональными элементами, равными нулю, а C — верхняя треугольная матрица с диагональными элементами, отличными от нуля, т. е. $\alpha = B + C$. Следовательно,

$$(E - B)x^{(k+1)} = Cx^{(k)} + \beta,$$

откуда

$$x^{(k+1)} = (E - B)^{-1}Cx^{(k)} + (E - B)^{-1}\beta.$$

Таким образом, метод Зейделя является методом простых итераций с матрицей правых частей $\alpha = (E - B)^{-1}$C и вектором правых частей $(E - B)^{-1}\beta$, и, следовательно, сходимость и погрешность метода Зейделя можно исследовать с помощью формул, выведенных для метода простых итераций, в которых вместо матрицы α подставлена матрица $(E - B)^{-1}$C, а вместо вектора правых частей β — вектор $(E - B)^{-1}\beta$. Для практических вычислений важно, что в качестве достаточных условий сходимости метода Зейделя могут быть использованы условия, приведенные выше для метода простых итераций ($\|\alpha\| < 1$ или, если используется эквивалентная СЛАУ в форме (2.32), — диагональное преобладание матрицы A). В случае выполнения этих

условий для оценки погрешности на k-ой итерации можно использовать выражение

$$\varepsilon^{(k)} = \frac{\|C\|}{1-\|\alpha\|}\left\|x^{(k)} - x^{(k-1)}\right\|.$$

Отметим, что, как и метод простых итераций, метод Зейделя может сходиться и при нарушении условия $\|\alpha\| < 1$. В этом случае $\varepsilon^{(k)} = \left\|x^{(k)} - x^{(k-1)}\right\|$.

Пример 2.8. Методом Зейделя решить СЛАУ из примера 2.7.

Решение.

Приведение СЛАУ к эквивалентному виду аналогично примеру (2.7). Диагональное преобладание элементов исходной матрицы СЛАУ гарантирует сходимость метода Зейделя.

Итерационный процесс выглядит следующим образом:

$$x^{(0)} = (\ 1{,}2 \quad 1{,}3 \quad 1{,}4\ )^T,$$

$$\begin{cases} x_1^{(1)} = 1{,}2 - 0{,}1 \cdot 1{,}3 - 0{,}1 \cdot 1{,}4 = 0{,}93, \\ x_2^{(1)} = 1{,}3 - 0{,}2 \cdot 0{,}93 - 0{,}1 \cdot 1{,}4 = 0{,}974, \\ x_3^{(1)} = 1{,}4 - 0{,}2 \cdot 0{,}93 - 0{,}2 \cdot 0{,}974 = 1{,}0192, \end{cases}$$

$$\begin{cases} x_1^{(2)} = 1{,}2 - 0{,}1 \cdot 0{,}974 - 0{,}1 \cdot 1{,}0192 = 1{,}0007, \\ x_2^{(2)} = 1{,}3 - 0{,}2 \cdot 1{,}0007 - 0{,}1 \cdot 1{,}0192 = 0{,}998, \\ x_3^{(2)} = 1{,}4 - 0{,}2 \cdot 1{,}0007 - 0{,}2 \cdot 0{,}998 = 1{,}0003. \end{cases}$$

Таким образом, уже на второй итерации погрешность $\left\|x^{(2)} - x^*\right\| < 10^{-2} = \varepsilon$, т. е. метод Зейделя, в данном случае сходится быстрее метода простых итераций.

2.1.8. Метод Зейделя для нормальных СЛАУ. Пусть дана СЛАУ

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}. \qquad (2.39)$$

Если СЛАУ имеет диагональное преобладание, т. е. $|a_{ii}| > \sum\limits_{\substack{j=1\\ i\neq j}}^{n} |a_{ij}|$, $i = \overline{1, n}$, методы простых итераций и Зейделя сходятся. Однако единственное решение СЛАУ (2.39) существует и тогда, когда нет диагонального преобладания, но матрица A является невырожденной (т. е. $\det A \neq 0$).

С помощью элементарных преобразований можно из невырожденной матрицы построить матрицу с диагональным преобладанием, для которой рассмотренные итерационные методы будут сходиться. Однако существует более простой способ, а именно нормализовать СЛАУ (2.39) с невырожденной матрицей A и применить к ней метод Зейделя.

СЛАУ называется *нормальной*, если ее матрица симметрична ($A = A^T$) и положительно определена. *Положительно определенная матрица* — матрица, для которой квадратичная форма (Ax, x)>0, или матрица, у которой все собственные значения положительны. По критерию Сильвестра для положительно определенной матрицы все диагональные миноры матрицы положительны:

$$a_{11} > 0, \qquad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \ldots, \quad \det A > 0.$$

Из любой невырожденной матрицы можно сделать нормальную, домножив ее на транспонированную слева. Умножим (2.39) слева на A^T, получим

$$A^T A x = A^T b, \quad \text{или} \quad Bx = c,$$

где $B = A^T A$ — нормальная матрица, $c + A^T b$.

Справедлива следующая теорема.

Теорема 2.3. *Метод Зейделя всегда сходится для нормальных СЛАУ.*

Пример 2.9. Для следующей СЛАУ применить один из итерационных методов решения:

$$\begin{cases} 2x_1 + 10x_2 + x_3 = 13, \\ 10x_1 + x_2 + x_3 = 12, \\ 2x_1 + 2x_2 + 10x_3 = 14. \end{cases}$$

Решение.

Эквивалентная СЛАУ и ее матрица имеют вид

$$\begin{cases} x_1 = 6{,}5 - 5x_2 - 0{,}5x_3, \\ x_2 = 12 - 10x_1 - x_3, \\ x_3 = 1{,}4 - 0{,}2x_1 - 0{,}2x_2, \end{cases} \qquad \alpha = \begin{pmatrix} 0 & -5 & -0{,}5 \\ -10 & 0 & -1 \\ -0{,}2 & -0{,}2 & 0 \end{pmatrix},$$

$\|\alpha\|_1 = 11 \gg 1$, т. е. не выполнено достаточное условие сходимости метода простых итераций. Нормализуем исходную СЛАУ: $A^T A x = A^T b$, где

$$A^T A = \begin{bmatrix} 108 & 34 & 32 \\ 34 & 105 & 31 \\ 32 & 31 & 102 \end{bmatrix}; \quad A^T b = \begin{bmatrix} 174 \\ 170 \\ 165 \end{bmatrix}.$$

Результирующая СЛАУ имеет симметрическую матрицу и диагональное преобладание. Следовательно, матрица эквивалентной системы по норме меньше единицы, что достаточно для сходимости метода Зейделя.

Отметим, однако, что рассмотренный способ нормализации существенно ухудшает обусловленность СЛАУ. Кроме того, его применение к разреженным матрицам может приводить к потере этого свойства. Поэтому при решении СЛАУ большой размерности данный подход используется редко.

УПРАЖНЕНИЯ.

2.21. Методом простых итераций с точностью $\varepsilon = 10^{-3}$ решить СЛАУ, приведя к эквивалентной системе и оценив предварительно число итераций:

$$\text{а)} \begin{cases} 12x_1 - 3x_2 + 2x_3 - x_4 = 8, \\ -x_1 + 6x_2 - x_3 + x_4 = 12, \\ 3x_1 + 2x_2 - 8x_3 + 2x_4 = -9, \\ 2x_1 - x_2 - x_3 + 5x_4 = 17; \end{cases}$$

$$\text{б)} \begin{cases} 6x_1 + 2x_2 - x_3 + x_4 = 0, \\ x_1 - 5x_2 - x_3 + 2x_4 = 4, \\ -3x_1 + 2x_2 + 8x_3 + x_4 = 9, \\ 2x_1 - x_2 + 2x_3 + 7x_4 = -7. \end{cases}$$

2.22. Методом Зейделя с точностью $\varepsilon = 10^{-3}$ решить задачи в задании 2.21 и сравнить по количеству итераций с методом простых итераций.

2.23. Методом Зейделя с точностью $\varepsilon = 10^{-3}$ решить СЛАУ, приведя их предварительно к виду, удобному для итераций:

$$\text{а)}\ \begin{cases} 2{,}7x_1 + 9{,}3x_2 + 1{,}3x_3 = 2{,}1, \\ 3{,}5x_1 + 1{,}7x_2 + 2{,}8x_3 = 1{,}7, \\ 4{,}1x_1 + 5{,}8x_2 - 1{,}7x_3 = 0{,}8; \end{cases}$$

$$\text{б)}\ \begin{cases} 1{,}7x_1 - 2{,}8x_2 + 1{,}9x_3 = 0{,}7, \\ 2{,}1x_1 + 3{,}4x_2 + 1{,}8x_3 = 1{,}1, \\ 4{,}2x_1 - 1{,}7x_2 + 1{,}3x_3 = 2{,}8. \end{cases}$$

2.24. Оценить количество итераций k в задачах задания 2.21 для различных значений точности: $\varepsilon_1 = 10^{-2}$; $\varepsilon_2 = 10^{-3}$; $\varepsilon_3 = 10^{-4}$; $\varepsilon_4 = 10^{-5}$. Построить график зависимости $k(\varepsilon)$.

2.25. Сходится ли метод простых итераций для задач задания 2.23, матрицы которых не имеют диагонального преобладания? Максимальное собственное значение оценить степенным методом (см. п. 2.4.3).

§ 2.2. Численные методы решения нелинейных и трансцендентных уравнений

Рассмотрим уравнение вида

$$f(x) = 0, \tag{2.40}$$

где $f(x)$ — любая нелинейная или трансцендентная функция, например $f(x) = \exp(\operatorname{tg} x) - x^3 \sin x$.

Для нахождения корней уравнения (2.40) различают следующие два этапа.

1. Отделение корней, т. е. нахождение таких интервалов по аргументу x, внутри каждого из которых существует только один корень уравнения (2.40).

2. Уточнение корней заключается в применении некоторого итерационного метода, в результате которого корень уравнения (2.40) может быть получен с любой наперед заданной точностью ε. При этом, останавливая процесс на какой-либо конечной

итерации, необходимо оценить погрешность по сравнению с точным корнем, который неизвестен.

2.2.1. Способы отделения корней. Наиболее употребимыми на практике способами отделения корней являются следующие.

1. Графический.

2. Метод половинного деления.

В графическом способе строится график функции $f(x)$ (рис.2.2) и приближенно определяются ее нули (или корни уравнения $f(x) = 0$) ξ_i, $i = 1, 2, 3, \ldots$ Заключив эти нули в интервалы

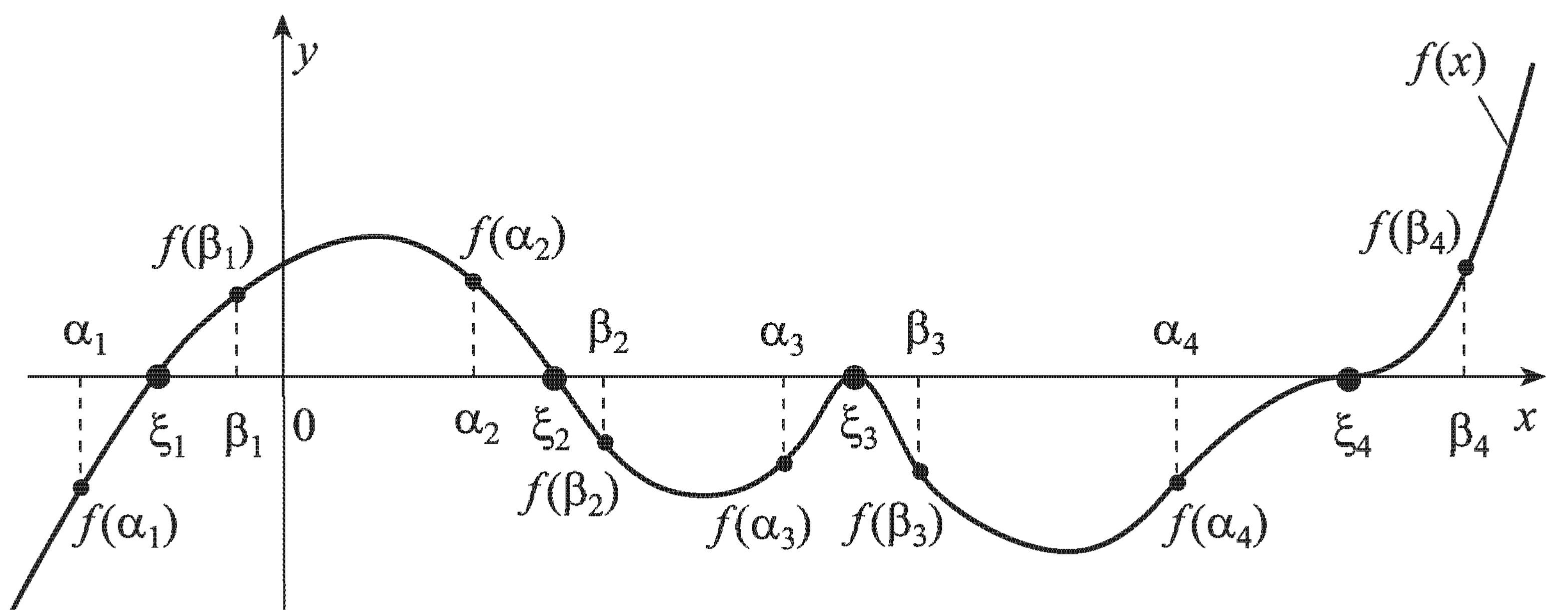


Рис. 2.2. Графический метод отделения корней

$\xi_i \in (\alpha_i, \beta_i)$, на границах которых выполняются условия $f(\alpha_i) \times$ $\times\, f(\beta_i) < 0$, $i = 1, 2, \ldots$, и знаки производных первого и второго порядков $f'(x)$, $f''(x)$ на интервалах (α_i, β_i) постоянны, можно утверждать, что внутри каждого из этих интервалов находится один корень уравнения (2.40).

Если при этом $f(\xi_i) = f'(\xi_i) = 0$, а $f''(\xi_i) \neq 0$, то корень ξ_i — двукратный корень (точка ξ_3 на рис.2.2); если $f(\xi_i) = f'(\xi_i) =$ $= f''(\xi_i) = 0$, а $f'''(\xi_i) \neq 0$, то ξ_i — трехкратный корень (точка ξ_4 на рис.2.2) и т. д.

В методе половинного деления область определения функции $f(x) : x \in [a, b]$ делят на 2, 4, 8, 16, 32, … интервала и для каждого из них анализируют знаки функции на концах интервала; если они противоположны, то внутри интервала находится не менее одного корня; если знак первой производной на интервале постоянный, т. е. $\mathrm{sign} f'(x)|_{x \in (\alpha_i, \beta_i)} = \mathrm{const}$ (при выполнении предыдущего условия), то внутри интервала находится точно

один корень. В данном параграфе рассматриваются однократные (простые) вещественные корни нелинейных уравнений (2.40).

Пример 2.10. Отделить корни уравнения

$$f(x) = x^2 - e^{-x/2} = 0, \ x \in [-3; 3].$$

Решение.

Уравнению $x^2 - e^{-x/2} = 0$ соответствует эквивалентное уравнение $x^2 = e^{-x/2}$. Если построить графики функций $y_1 = x^2$ и $y_2 = e^{-x/2}$, то абсциссы точек пересечения графиков этих

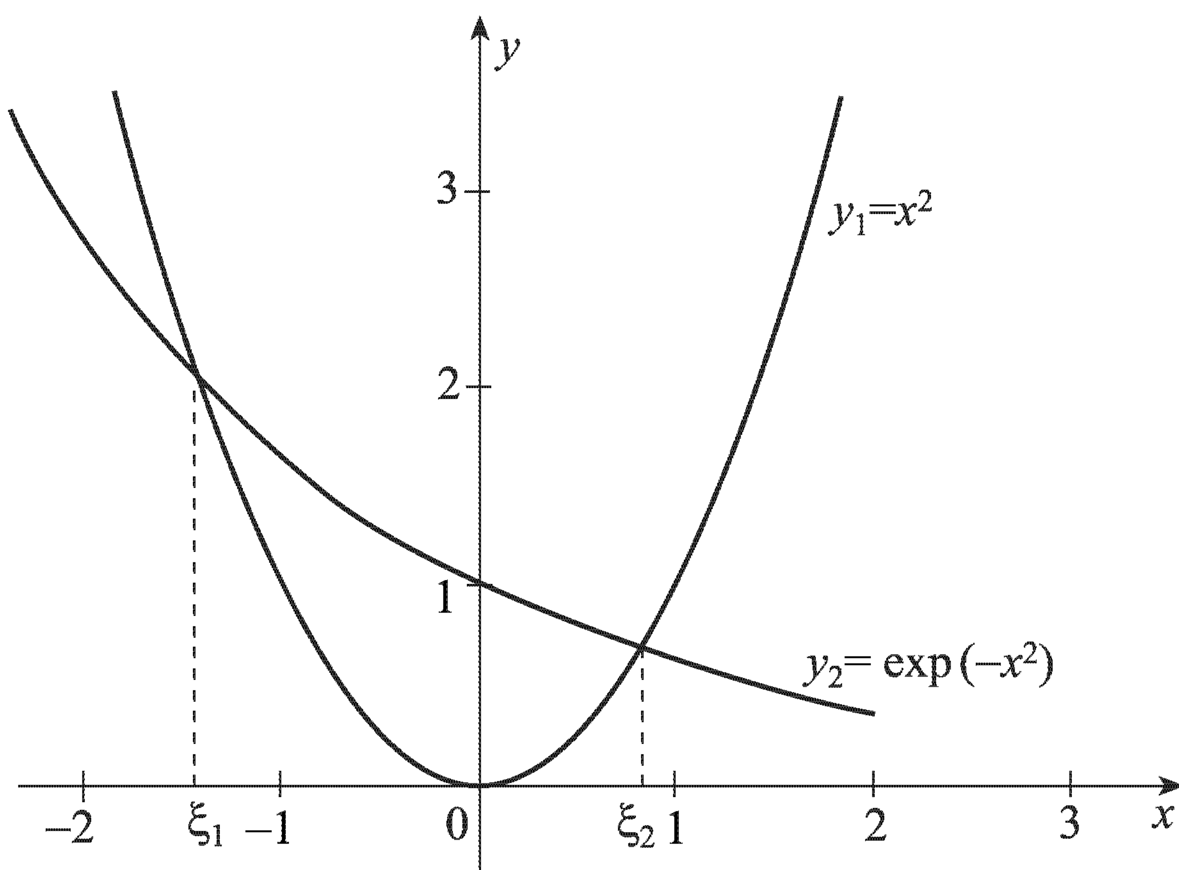


Рис. 2.3. Отделение корней

функций дадут искомые корни данного уравнения. Из рис. 2.3 видно, что $\xi_1 \in (-2;\ -1)$; $\xi_2 \in (0;\ 1)$.

2.2.2. Методы уточнения корней. Ниже рассматриваются следующие итерационные методы уточнения корней нелинейного уравнения (2.40).

1. Метод половинного деления.
2. Метод Ньютона (касательных).
3. Метод секущих (хорд).
4. Метод простых итераций.

Все эти методы являются итерационными.

Предположим, что $\overline{x}$ — приближенное значение корня, ξ — его точное значение. Возникает вопрос, какова погрешность $\xi - \overline{x}$ приближенного значения корня $\overline{x}$ по сравнению с его точным значением ξ, если последний неизвестен?

Для этого построим *невязку* $f(\xi) - f(\overline{x}) = -f(\overline{x})$, т. к. $f(\xi) = 0$. Применим к невязке теорему Лагранжа о конечных приращениях:

$$f(\xi) - f(\overline{x}) = f'(\alpha)(\xi - \overline{x}), \quad \alpha \in (\xi, \overline{x}),$$

откуда

$$\xi - \overline{x} = -\frac{f(\overline{x})}{f'(\alpha)}.$$

Так как точное значение α неизвестно, эту погрешность заменяют верхней оценкой:

$$|\xi - \overline{x}| \leqslant \frac{|f(\overline{x})|}{\min\limits_{x \in [\alpha_i, \beta_i]} |f'(x)|}. \tag{2.41}$$

Оценка погрешности (2.41) является довольно грубой. Поэтому в каждом итерационном методе уточнения корней, в силу ограничений применения метода, можно вывести свою оценку погрешности.

1. Метод половинного деления. Пусть $f(x) = 0$ и точное значение корня $\xi \in (a, b)$.

В методе половинного деления функция $f(x)$, $x \in [a, b]$, должна удовлетворять следующим двум условиям:

а) $f(a)f(b) < 0$;

б) $f(x)$ непрерывна на отрезке $x \in [a, b]$.

Метод половинного деления состоит в построении последовательности вложенных отрезков $[a_k, b_k] \subset [a_{k-1}, b_{k-1}] \subset \ldots \ldots \subset [a_0, b_0] \equiv [a, b]$, на концах которых удовлетворяются условия $f(a_k) f(b_k) < 0$, $k = 0, 1, 2, \ldots$ (рис. 2.4).

Опишем один шаг итерационного метода половинного деления. Пусть на $(k-1)$-м шаге найден отрезок $[a_{k-1}, b_{k-1}] \subset [a_0, b_0]$, на котором выполнено условие $f(a_{k-1}) f(b_{k-1}) < 0$. Этот отрезок делится пополам точкой $x_k = (a_{k-1} + b_{k-1})/2$, и вычисляется $f(x_k)$. Если $f(x_k) = 0$, то $x_k = (a_{k-1} + b_{k-1})/2$ — корень уравнения (2.40). Если $f(x_k) \neq 0$, то из двух половин отрезка выбирается та, на концах которой функция имеет противоположные знаки, так как корень находится внутри именно этой половины. Таким образом, $a_k = a_{k-1}$, $b_k = x_k$, если $f(a_{k-1}) f(x_k) < 0$; $a_k = x_k$, $b_k = b_{k-1}$, если $f(x_k) f(b_{k-1}) < 0$.

Если необходимо найти корень с точностью ε, то деление пополам продолжается до тех пор, пока не выполнится условие

$$b_k - a_k = (1/2)^k (b-a) \leqslant \varepsilon. \tag{2.42}$$

Тогда значение $\xi = (a_k + b_k)/2$ приближенно определяет корень с точностью ε.

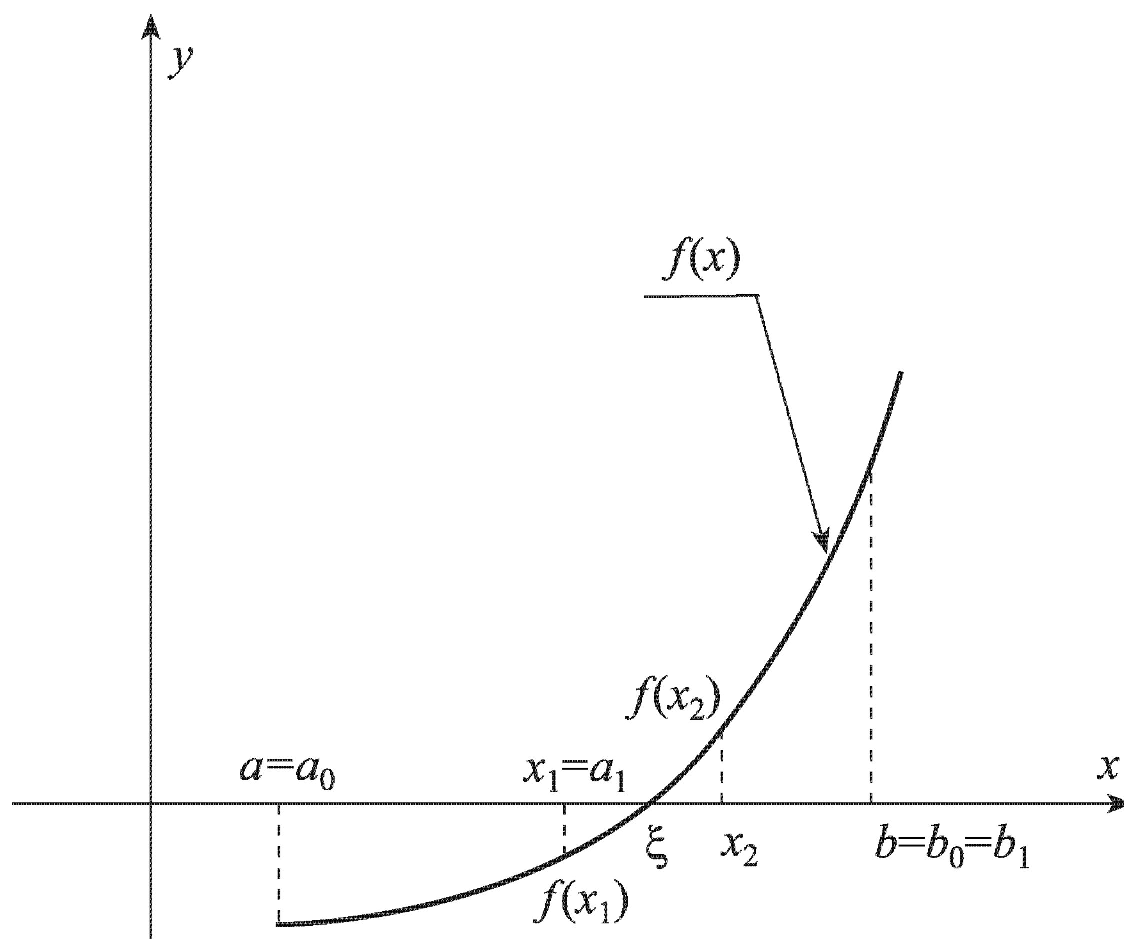


Рис. 2.4. К методу половинного деления

Из (2.42) можно оценить число k итераций, необходимых для достижения заданной точности ε: $k \geqslant \log_2((b-a)/\varepsilon)$. Отсюда видно, что для получения трех верных знаков $\left(\varepsilon = 10^{-3}\right)$ необходимо сделать ~ 10 итераций.

Покажем, что процесс итераций по методу половинного деления сходится, и сходится к корню уравнения (2.40). Для этого на основании (2.42) запишем погрешность, которая получится при остановке итерационного процесса на k-й итерации:

$$\xi - x_k = \frac{1}{2^k}(b-a). \tag{2.43}$$

Пусть $f(a_k)f(b_k) < 0$. Из (2.43) видно, что $\lim\limits_{k\to\infty}(\xi - x_k) = \lim\limits_{k\to\infty} \frac{1}{2^k}(b-a) = 0$, откуда следует $\lim\limits_{k\to\infty} x_k = \xi$, где ξ — точный корень.

Возьмем предел от произведения $f(a_k)\cdot f(b_k)$ при $k \to \infty$. Поскольку функция $f(x)$ непрерывна на $x \in [a,b]$ по условию

задачи, то знак предела можно внести под знак функции, т. е.

$$\lim_{k\to\infty} f(a_k)f(b_k) = \\ = f(\lim_{k\to\infty} a_k)f(\lim_{k\to\infty} b_k) = f(\xi)f(\xi) = f^2(\xi) < 0, \quad (2.44)$$

что невозможно и, следовательно, $f^2(\xi) = 0$, или $f(\xi) = 0$, т. е. ξ является единственным корнем уравнения (2.40).

При выводе (2.44) были использованы пределы $\lim\limits_{k\to\infty} a_k = \xi$ и $\lim\limits_{k\to\infty} b_k = \xi$. Действительно, последовательность $\{a_k\}$ является неубывающей, ограниченной сверху значением ξ, а последовательность $\{b_k\}$ является невозрастающей, ограниченной снизу значением ξ. Таким образом, пределом слева и справа является точный корень ξ.

***Достоинства*:**

а) метод половинного деления прост в алгоритмизации и программировании;

б) на функцию $f(x)$ не накладывается никаких ограничений, кроме требования непрерывности.

***Недостаток*:** метод очень медленно сходится, т. е. необходимо использовать большое число итераций для достижения заданной точности ε.

Априорная оценка количества итераций в соответствии с неравенством $k \geqslant \log_2[(b-a)/\varepsilon]$ довольно завышена, поэтому на практике итерационный процесс останавливается при выполнении неравенства

$$|x_{k+1} - x_k| \leqslant \varepsilon, \quad \xi \approx \frac{x_{k+1} + x_k}{2}.$$

Пример 2.11. Методом половинного деления уточнить наибольший корень уравнения $x^2 - e^{-x} = 0$ с точностью $\varepsilon = 10^{-2}$.

Решение.

Выделим наибольший корень: $a = 0{,}5$; $b = 1{,}0$. $\xi \in (0{,}5;\, 1{,}0)$.

1) $a_0 \equiv a = 0{,}5$; $b_0 \equiv b = 1{,}0$; $x_1 = (a_0 + b_0)/2 = 0{,}75$; $f(x_1) = 0{,}0901 > 0$; $f(a_0)f(x_1) = -0{,}3565 \cdot 0{,}0901 < 0$; $\xi \in (a_0, x_1)$; тогда $a_1 = a_0 = 0{,}5$; $b_1 = x_1 = 0{,}75$; $|b_1 - a_1| = 0{,}25 > \varepsilon$.

Продолжая этот процесс, получаем на 6-й итерации: $a_6 = 0{,}7032$; $b_6 = 0{,}711$; $|b_6 - a_6| = 0{,}0078 < \varepsilon$; $\xi = (b_6 + a_6)/2 = 0{,}7071$.

УПРАЖНЕНИЯ.

Методом половинного деления с точностью $\varepsilon = 10^{-2}$ уточнить корни следующих уравнений:

2.26. $\operatorname{tg}(1{,}9x) - 2{,}8x = 0$;

2.27. $\ln(8x) = 9x - 3$;

2.28. $\sin(2{,}2x) - x = 0$;

2.29. $\exp(-0{,}8x) - 4x = 0$;

2.30. $0{,}89x^3 - 2{,}8x^2 - 3{,}7x + 11{,}2 = 0$.

2. Метод Ньютона уточнения корней. Пусть для уравнения (2.40) на интервале $x \in (a, b)$ отделен корень ξ.

В методе Ньютона функция $f(x)$ должна удовлетворять на отрезке $x \in [a, b]$ следующим условиям:

1) существование производных 1-го и 2-го порядков;

2) $f'(x) \neq 0$;

3) производные 1-го и 2-го порядков знакопостоянны ($\operatorname{sign}\{f'(x), f''(x)\} = \mathrm{const}$) на отрезке $x \in [a, b]$.

Пусть имеется значение корня на k-й итерации — x_k. Тогда значение корня на $(k+1)$-й итерации вычисляется следующим образом:

$$x_{k+1} = x_k + h_k, \tag{2.45}$$

где h_k — шаг, который подлежит определению.

Чтобы определить h_k, подставим x_{k+1} в функцию $f(x)$ и разложим ее в ряд Тейлора до третьего слагаемого включительно в окрестности точки x_k, получим

$$f(x_{k+1}) = f(x_k + h_k) = f(x_k) + f'(x_k)\,h_k + f''(\alpha)\,\frac{h_k^2}{2},$$
$$\alpha \in (x_k, x_{k+1}).$$

Положим в этом разложении линейную относительно h_k часть равной нулю, в результате чего находим значение шага h_k:

$$h_k = -\frac{f(x_k)}{f'(x_k)},$$

подставляя которое в (2.45), получаем

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots \tag{2.46}$$

За начальное приближение x_0 принимается один из концов отрезка $[a, b]$, а именно

$$x_0 = \begin{cases} a, & \text{если } f(a) \cdot f''(a) > 0; \\ b, & \text{если } f(b) \cdot f''(b) > 0. \end{cases} \tag{2.47}$$

Выражения (2.46), (2.47) называют *итерационным методом Ньютона* (*или касательных*) уточнения корней нелинейного уравнения (2.40).

Чтобы не вычислять значение первой производной $f'(x_k)$ на каждой итерации, достаточно вычислить ее в точке x_0 и полученное значение подставить в выражение (2.46), получим

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}, \quad k = 0, 1, 2, \ldots \tag{2.48}$$

Метод (2.47), (2.48) называют *модифицированным методом Ньютона*. Он обладает неоспоримым преимуществом перед методом (2.46), (2.47), но сходится медленнее.

Справедлива следующая теорема.

Теорема 2.4. (*достаточные условия сходимости метода Ньютона*). *Пусть* $f(x)$ *определена и дважды дифференцируема на отрезке* $x \in [a, b]$, *причем* $f(a) \cdot f(b) < 0$, *производные* $f'(x)$ *и* $f''(x)$ *знакопостоянны и* $f'(x) \neq 0$. *Тогда исходя из начального приближения* $x_0 \in [a, b]$, *удовлетворяющего неравенству* $f(x_0) \cdot f''(x_0) > 0$, *можно построить последовательность* (2.46), *сходящуюся к единственному корню* ξ *уравнения* (2.40) *на отрезке* $x \in [a, b]$ *с погрешностью, оцениваемой неравенством*

$$|\xi - x_{k+1}| \leqslant \frac{h_k^2}{2} \frac{M_2}{m_1}, \tag{2.49}$$
$$M_2 = \max_{x \in [a,b]} |f''(x)|, \quad m_1 = \min_{x \in [a,b]} |f'(x)|, \quad x \in [a, b].$$

Действительно, поскольку $f(x)$ непрерывна, то взяв предел от (2.46) при $k \to \infty$ и поменяв местами знаки предела и функции, получим

$$\lim_{k \to \infty} x_{k+1} = \lim_{k \to \infty} x_k - \frac{f\left(\lim\limits_{k \to \infty} x_k\right)}{f'\left(\lim\limits_{k \to \infty} x_k\right)}.$$

Поскольку последовательность $\{x_k\}$ — невозрастающая (неубывающая) числовая последовательность, ограниченная снизу

(сверху), то она сходится, т. е. существует предел $\lim\limits_{k\to\infty} x_k = \xi$. Тогда из последнего равенства ($f'(\xi) \neq 0$ по условию) имеем

$$\xi = \xi - \frac{f(\xi)}{f'(\xi)}, \quad \text{откуда} \quad f(\xi) = 0,$$

т. е. предельное значение ξ является корнем уравнения (2.40).

Поскольку невязка равна $f(x_{k+1}) - f(\xi) = f(x_{k+1})$, то из разложения $f(x_{k+1})$ в ряд Тейлора имеем выражение для невязки в виде

$$f(x_{k+1}) = f''(\alpha)\frac{h_k^2}{2},$$

подставляя которое в общую погрешность приближенных методов (2.41) при $\overline{x}=x_{k+1}$, получаем верхнюю оценку погрешности в виде (2.49), где $h_k = x_{k+1} - x_k$.

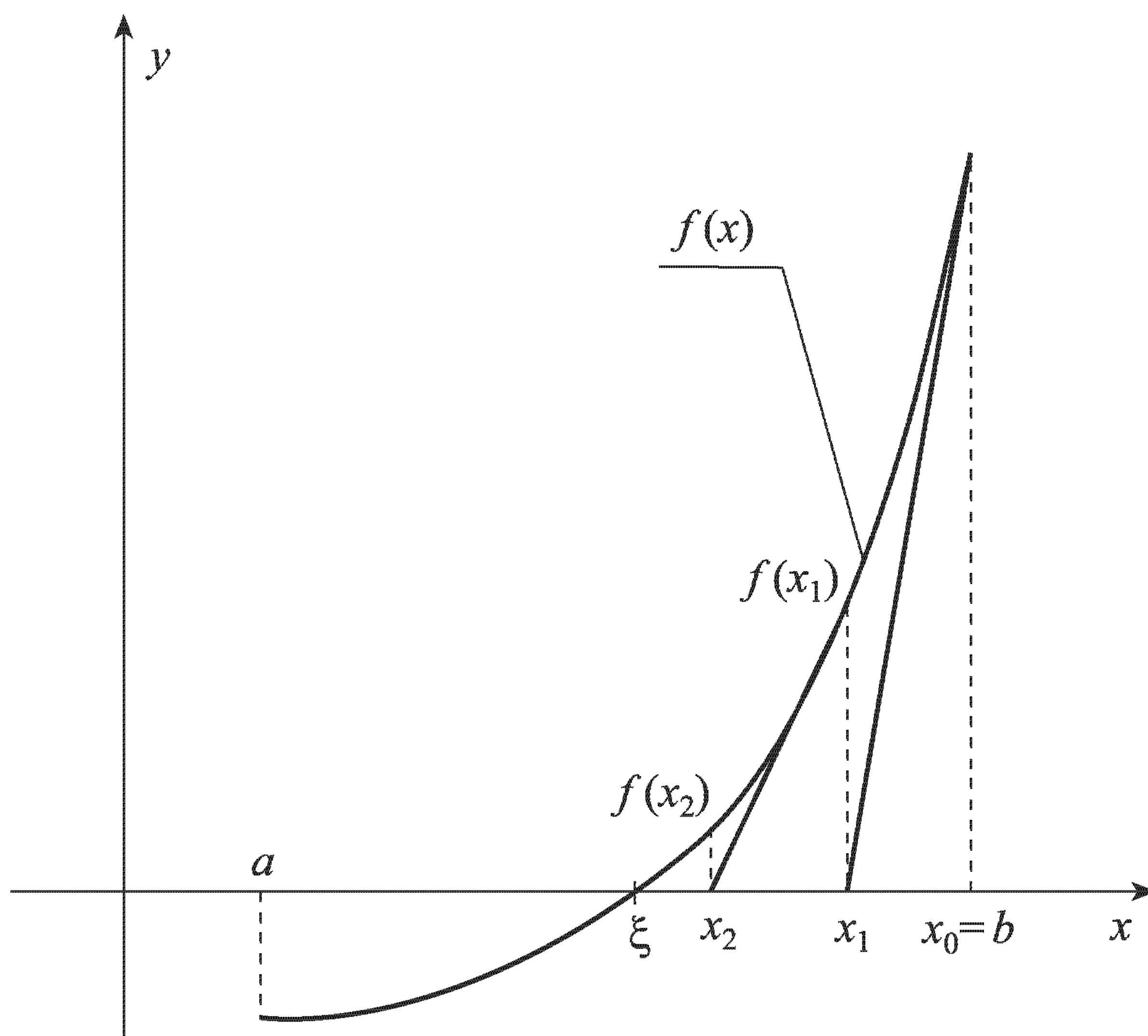


Рис. 2.5. К методу Ньютона

Для случая, приведенного на рис. 2.5, за начальное приближение принимается $x_0 = b$, так как $f(b) \cdot f''(b) > 0$.

На основе (2.49) можно записать, что $|\xi - x_{k+1}| \leqslant (M_2/2m_1)\,|\xi - x_k|^2$, и если $M_2/2m_1 < 1$ и $|\xi - x_k| < 10^{-m}$, то $|\xi - x_{k+1}| < 10^{-2m}$, т. е. когда приближение x_k

имеет m верных знаков, x_{k+1} будет иметь не менее $2m$ верных знаков, т. е. метод Ньютона имеет квадратичную сходимость.

Поскольку верхняя оценка (2.49) сложна для вычисления, на практике итерационный процесс останавливают при выполнении условия $|x_{k+1} - x_k| \leqslant \varepsilon$, $\xi \approx x_{k+1}$, где ε — заданная точность.

Метод Ньютона называют еще методом касательных, поскольку в этом методе на каждой итерации к графику функции $f(x)$ проводится касательная в точке $(x_k, f(x_k))$ до пересечения с осью абсцисс (рис. 2.5).

Пример 2.12. Методом Ньютона с точностью $\varepsilon = 10^{-2}$ уточнить наибольший корень уравнения $f(x) = x^2 - e^{-x} = 0$, $x \in (0,5;\ 1,0)$.

Решение.

Начальное приближение $x_0 = 1,0$, так как $f(1,0) > 0$; $f''(1,0) > 0$. Таким образом $x_0 = 1,0$; $x_1 = 1 - f(1)/f'(1) = 0,733$; $x_2 = 0,733 - f(0,733)/f'(0,733) = 0,7038$; $|x_2 - x_1| = 0,092 > \varepsilon$; $x_3 = x_2 - f(x_2)/f'(x_2) = 0,7034$; $|x_3 - x_2| = 0,3414 \cdot 10^{-3} < \varepsilon$, т. е. $\xi \approx 0{,}7034$.

УПРАЖНЕНИЯ.

Методом Ньютона с точностью $\varepsilon = 10^{-3}$ уточнить корни следующих уравнений:

2.31. $\mathrm{tg}\,(1,262x) - 1,84x = 0$;

2.32. $\ln(3,66x) = 4,12x - 1,5$;

2.33. $2,33\sin(2,86x) = 2x$;

2.34. $0,7\exp(-0,59x) - x = 0$;

2.35. $1,2x^3 - 3,53x^2 - 1,36x + 7,11 = 0$.

3. Метод секущих (хорд). Заменяя в алгоритме Ньютона производную $f'(x_k)$ приближенно отношением конечных разностей:

$$f'(x_k) \approx \frac{f(b) - f(x_k)}{b - x_k},$$

получим алгоритм метода хорд с неподвижным правым концом (рис. 2.6 *а*):

$$x_{k+1} = x_k - \frac{f(x_k)}{f(b) - f(x_k)}(b - x_k), \quad x_0 = a, \tag{2.50}$$

или с неподвижным левым концом (рис. 2.6 *б*):

$$x_{k+1} = x_k - \frac{f(x_k)}{f(a) - f(x_k)}(a - x_k), \quad x_0 = b. \tag{2.51}$$

В качестве неподвижного конца отрезка принимается граница $x = b$ (формула (2.50)), если $f(b) \cdot f''(b) > 0$; тогда $x_0 = a$, или $x = a$, если $f(a) \cdot f''(a) > 0$, тогда $x_0 = b$ (формула (2.51)).

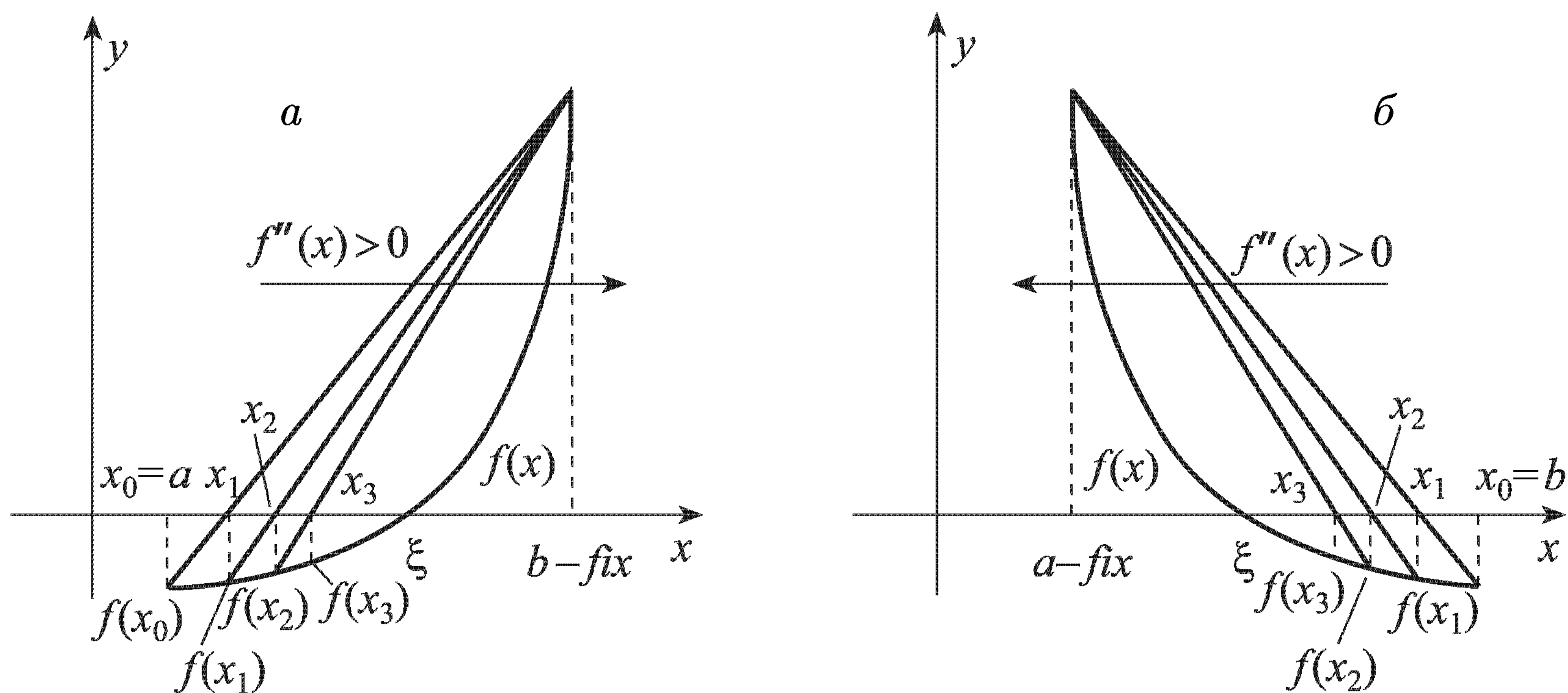


Рис. 2.6. К методу секущих (стрелки показывают направление сходимости)

Сходимость и погрешность метода хорд определяется так же, как и в методе Ньютона. На рис. 2.6 стрелками показано направление сходимости.

4. Метод простых итераций. Метод простых итераций уточнения корней уравнения (2.40) состоит в замене этого уравнения эквивалентным ему уравнением

$$x = \varphi(x), \quad x \in (a, b) \tag{2.52}$$

и построении последовательности

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, 2, \ldots, \tag{2.53}$$

где $x_0 \in (a, b)$, например $x_0 = (a + b)/2$.

Если не удается выразить x из уравнения (2.40), то эквивалентное уравнение и эквивалентную функцию можно построить, например, так:

$$x = x + f(x), \quad \varphi(x) = x + f(x).$$

Последовательность (2.53) называют *методом простых итераций* уточнения корней уравнения (2.40).

Для ответа на вопросы: сходится ли последовательность (2.53)? и, если сходится, является ли предельное значение корнем уравнения (2.52), а следовательно, и уравнения (2.40)? имеет место следующая теорема.

Теорема 2.5 (*достаточные условия сходимости метода простых итераций*). *Пусть функция* $\varphi(x)$ *в эквивалентном уравнении* (2.52) *определена и дифференцируема на отрезке* $x \in \in [a, b]$. *Тогда, если существует число* q *такое, что*

$$\left|\varphi'(x)\right| \leqslant q < 1 \tag{2.54}$$

на отрезке $[a, b]$, *то последовательность* (2.53) *сходится к единственному корню уравнения* (2.52) *при любом начальном приближении* $x_0 \in [a, b]$.

Доказательство. Запишем итерационный процесс (2.53) на двух соседних итерациях:

$$\begin{cases} x_k = \varphi(x_{k-1}), \\ x_{k+1} = \varphi(x_k). \end{cases}$$

Вычитая из второго выражения первое, получим

$$x_{k+1} - x_k = \varphi(x_k) - \varphi(x_{k-1}). \tag{2.55}$$

Применим к правой части выражения (2.55) теорему Лагранжа о конечных приращениях, поскольку по условию $\varphi(x)$ — дифференцируемая функция:

$$x_{k+1} - x_k = \varphi'(\alpha)(x_k - x_{k-1}), \quad \alpha \in (x_{k-1}, x_k).$$

Используя условия (2.54) теоремы, из последнего равенства получаем неравенство

$$|x_{k+1} - x_k| \leqslant q\,|x_k - x_{k-1}|\,,$$

на основе которого можно выписать следующую цепочку неравенств:

$$\begin{aligned} &|x_2 - x_1| \leqslant q\,|x_1 - x_0|\,, \\ &|x_3 - x_2| \leqslant q\,|x_2 - x_1| \leqslant q^2\,|x_1 - x_0|\,, \\ &|x_4 - x_3| \leqslant q^3\,|x_1 - x_0|\,, \\ &\dots\dots\dots\dots\dots\dots\dots\dots \\ &|x_{k+1} - x_k| \leqslant q^k\,|x_1 - x_0|\,. \end{aligned} \tag{2.56}$$

Тогда на основе тождества

$$x_{k+1} \equiv x_0 + (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_{k+1} - x_k)$$

и неравенства

$$|x_{k+1}| \leqslant |x_0| + |x_1 - x_0| + |x_2 - x_1| + \ldots + |x_{k+1} - x_k|$$

с использованием в нем цепочки (2.56), получим следующую оценку:

$$|x_{k+1}| \leqslant |x_0| + \left(1 + q + q^2 + \ldots + q^k\right) |x_1 - x_0|, \tag{2.57}$$

в круглых скобках которой находится сумма геометрической прогрессии. Взяв предел от (2.57) при $k \to \infty$, получим выражение

$$\lim_{k\to\infty} |x_{k+1}| \leqslant |x_0| + |x_1 - x_0| \lim_{k\to\infty} \sum_{n=0}^{k} q^n,$$

в котором сумма ряда геометрической прогрессии равна $(1-q)^{-1}$, поскольку по условию теоремы $|q| < 1$.

Таким образом,

$$\lim_{k\to\infty} |x_{k+1}| \leqslant x_0 + \frac{|x_1 - x_0|}{1-q} < M < \infty.$$

То есть существует конечный предел последовательности (2.53). Обозначим его через ξ, т. е.

$$\lim_{k\to\infty} x_{k+1} = \xi.$$

Для ответа на 2-й вопрос о том, является ли предельное значение ξ корнем уравнения (2.52), перейдем в (2.53) к пределу при $k \to \infty$ и, поскольку функция $\varphi(x)$ дифференцируема и, следовательно, непрерывна, знаки предела и функции можно менять местами; получим

$$\lim_{k\to\infty} x_{k+1} = \varphi\left(\lim_{k\to\infty} x_k\right),$$

откуда следует равенство

$$\xi = \varphi(\xi),$$

то есть ξ является единственным корнем уравнения (2.52), а следовательно, и уравнения (2.40).

При ответе на 3-й вопрос о погрешности метода можно показать, что если ξ — неизвестный точный корень уравнения (2.52) и итерационный процесс остановлен на k-й итерации, то погрешность, которая допускается этим приближением, оценивается

сверху выражением [1]

$$|\xi - x_k| < \frac{q}{1-q}\,|x_k - x_{k-1}| \tag{2.58}$$

или выражением

$$|\xi - x_k| < \frac{q^k}{1-q}\,|x_1 - x_0|\,. \tag{2.59}$$

Если задана точность ε, то итерации можно остановить в случае выполнения условия, полученного из (2.58):

$$|x_k - x_{k-1}| < \frac{1-q}{q}\varepsilon,$$

или из выражения (2.59) оценить минимальное количество итераций для достижения заданной точности ε:

$$k > \frac{1}{\ln q}\ln\frac{\varepsilon(1-q)}{|x_1 - x_0|}. \tag{2.60}$$

Вывод выражения (2.58) или (2.59) аналогичен выводу выражения (2.38) в методе простых итераций для СЛАУ (п. 2.1.6), если вместо $\|\alpha\|$ использовать значение q.

Неравенство (2.60) дает завышенное число итераций, поэтому на практике итерационный процесс останавливают при выполнении условия

$$|x_{k+1} - x_k| \leqslant \varepsilon, \quad \xi \approx x_{k+1}.$$

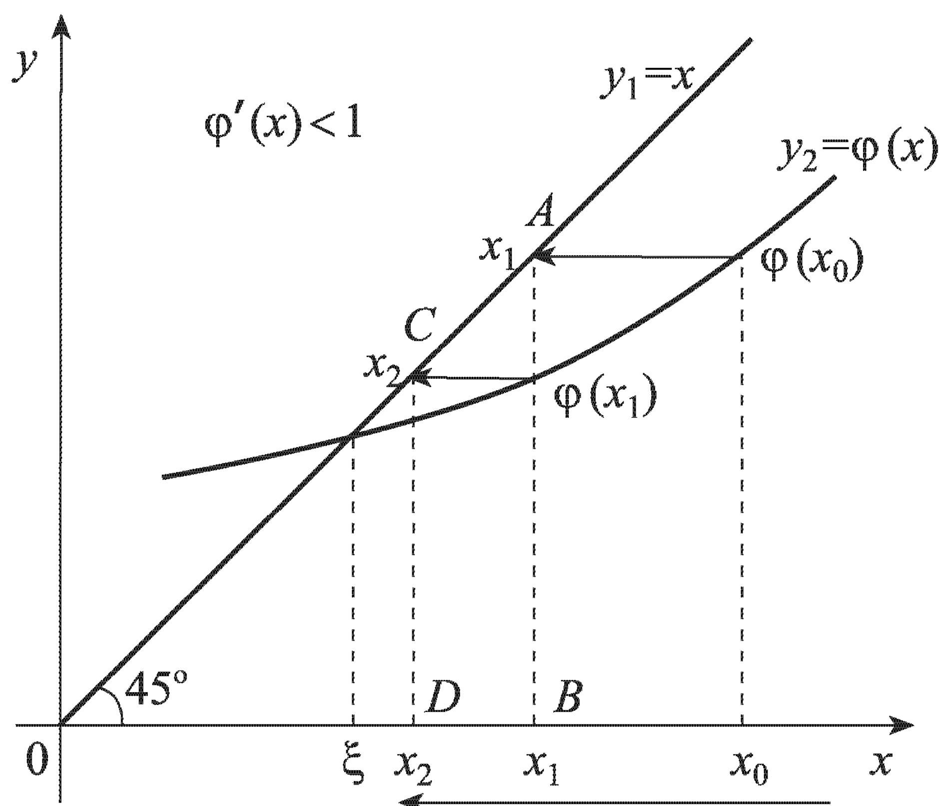


Рис. 2.7. К методу простых итераций в случае $\varphi'(x) < 1$

Геометрическая интерпретация метода простых итераций. Из рис. 2.7 видно, что $|\varphi'(x)| < 1$, так как тангенс

угла наклона касательной к графику функции $y_2 = \varphi(x)$ меньше $\mathrm{tg}(45^\circ) = 1$. Следовательно, для произвольного начального приближения x_0 в соответствии с 1-й итерацией в (2.53) при $k = 0$ определяется $\varphi(x_0)$, которое равно значению x_1 на графике функции $y_1 = x$, а поскольку треугольник OAB прямоугольный и равнобедренный, то $OB = x_1$. На второй итерации в (2.53) при $k = 1$ определяется $\varphi(x_1)$, которое равно значению x_2 на графике функции $y_1 = x$, а поскольку треугольник OCD — равнобедренный и прямоугольный, то $CD = OD = x_2$, т. е. итерационные значения x_0, x_1, x_2, ... стремятся в сторону точного корня ξ (указано стрелкой справа налево).

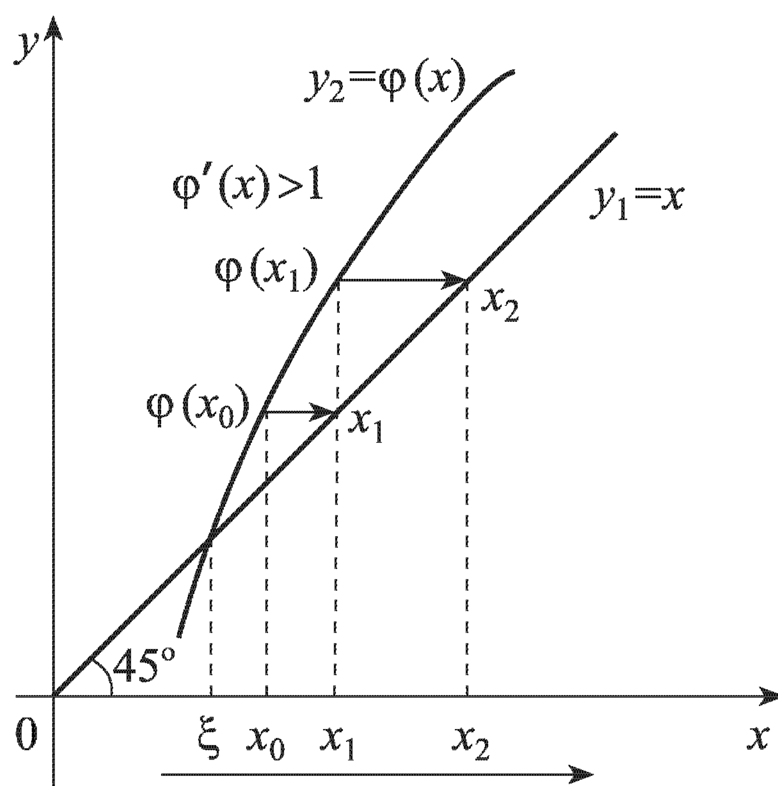


Рис. 2.8. К методу простых итераций в случае $\varphi'(x) > 1$

На рис. 2.8 $|\varphi'(x)| > 1$. Из рисунка видно, что итерационный процесс расходится (приближения корня x_0, x_1, x_2, ... стремятся от корня ξ).

На рис. 2.9 представлен случай $|\varphi'(x)| < 1$, $\varphi'(x) < 0$. Процесс итераций сходится с двух сторон, т. е. приближения корня находятся то слева, то справа от точного корня ξ.

Итерационный процесс (2.53) можно использовать и в случае, когда в эквивалентном уравнении $|\varphi(x)| > 1$. Для этого вернемся к исходному уравнению (2.40) и построим эквивалентное уравнение в виде

$$x = x \pm \frac{f(x)}{\max\limits_{x\in[a,b]} |f'(x)|},$$

где берется знак минус, если $f'(x) > 0$, и плюс, если $f'(x) < 0$.

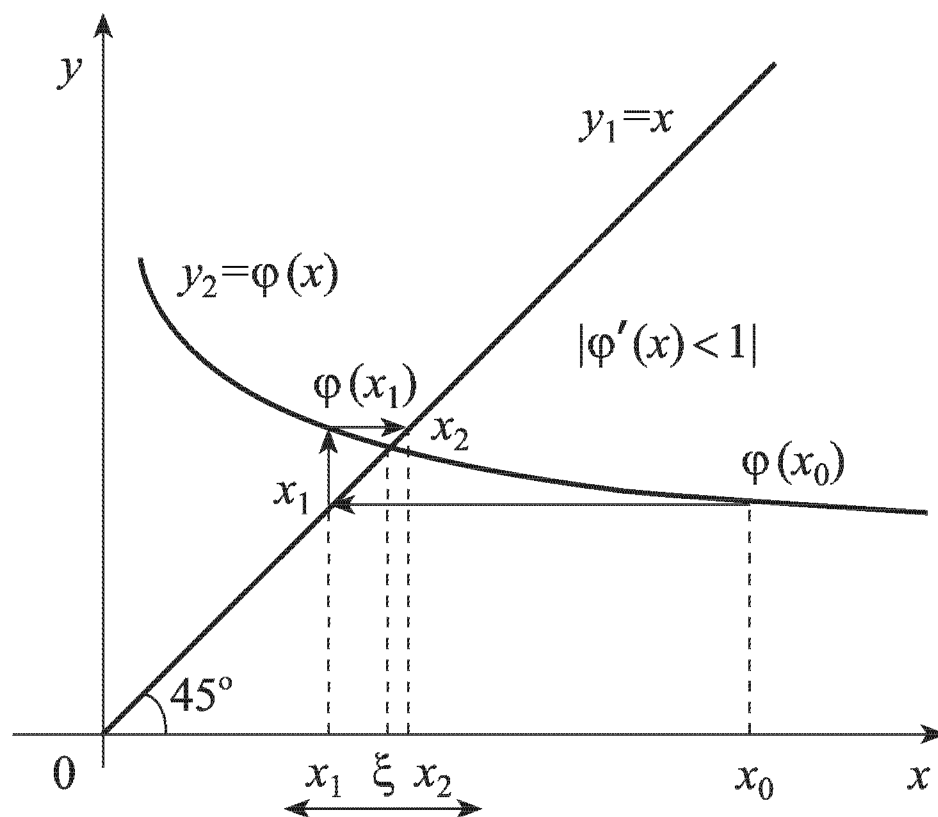


Рис. 2.9. К методу простых итераций в случае $|\varphi'(x)| < 1$, $\varphi'(x) < 0$

Тогда в качестве эквивалентной функции $\varphi(x)$ можно принять функцию

$$\varphi(x) = x \pm \frac{f(x)}{\max\limits_{x\in[a,b]} |f'(x)|},$$

для которой

$$\varphi'(x) = 1 \pm \frac{f'(x)}{\max\limits_{x\in[a,b]} |f'(x)|} < 1.$$

Пример 2.13. Методом простых итераций с точностью $\varepsilon = 10^{-2}$ уточнить наибольший корень уравнения $f(x) = x^2 - e^{-x} = 0$, $x \in [0{,}5;\ 1{,}0]$.

Решение.

Эквивалентное уравнение

$$x = \exp(-x/2);$$

$$\varphi(x) = \exp(-x/2);\ q = \max_{x\in[0{,}5;\,1{,}0]} \left|\varphi'(x)\right| = 0{,}3894 < 1;$$

$$x_0 = (0{,}5 + 1{,}0)/2 = 0{,}75; \quad x_1 = \exp(-0{,}375) = 0{,}687;$$

$$x_2 = \exp(-0{,}3437) = 0{,}7091;$$

$$|x_2 - x_1| = 0{,}02184 > (1 - q)\cdot 10^{-2}/q = 0{,}015;$$

$$x_3 = \exp(-0{,}3546) = 0{,}7015; \quad |x_3 - x_2| = 0{,}007646 < 0{,}015;$$

$\xi \approx 0{,}7015.$

УПРАЖНЕНИЯ.

Методом простых итераций с точностью $\varepsilon = 10^{-2}$ уточнить корни следующих уравнений:

2.36. $\operatorname{tg}(2{,}2x) - 3{,}2x = 0$;

2.37. $\ln(4{,}6x) = 5{,}2x - 1{,}5$;

2.38. $5{,}6\sin(4{,}8x) - 4{,}5x = 0$;

2.39. $0{,}8\exp(-0{,}6x) - x = 0$;

2.40. $0{,}17x^3 - 0{,}57x^2 - 1{,}6x + 3{,}7 = 0$.

2.2.3. Скорость сходимости. Процедура Эйткена ускорения сходимости. Скорость сходимости — одна из важнейших характеристик итерационных методов нахождения корней уравнения (2.40) [5].

Определение 1. *Говорят, что итерационный метод сходится со скоростью геометрической прогрессии, знаменатель которой* $q < 1$, *если справедлива оценка*

$$|\xi - x_k| \leqslant c_0 q^k, \quad k = 0, 1, 2, \ldots \tag{2.61}$$

Определение 2. *Число* p $(p \geqslant 1)$ *называют порядком сходимости метода, если в области сходимости имеет место оценка*

$$|\xi - x_{k+1}| \leqslant c \cdot |\xi - x_k|^p, \quad c > 0.$$

Если $p = 1$ и $c < 1$, то метод обладает *линейной скоростью сходимости*, при $p = 2$ — *квадратичной*, $p = 3$ — *кубичной* и т. д.

Имеют место следующие утверждения.

Лемма 2.2. *Если итерационный метод обладает линейной скоростью сходимости (порядок сходимости* $p = 1$*), то этот метод сходится со скоростью геометрической прогрессии со знаменателем* $q = c$ *и* $c_0 = \xi - x_0$ *и имеет место оценка погрешности* $|\xi - x_k| \leqslant q^k\,|\xi - x_0|$.

Лемма 2.3. *Если итерационный метод обладает* p*-м порядком сходимости (сверхлинейная скорость сходимости при* $p > 1$*), то этот метод сходится со скоростью показательной функции, имеющей основание* q *и показатель степени* p^k*, причем справедлива следующая оценка:* $|\xi - x_k| \leqslant c \cdot q^{p^k}$, $q = c^{1/(p-1)} \cdot |\xi - x_0|$.

Таким образом, метод половинного деления обладает линейной скоростью сходимости ($p = 1$), и, следовательно, сходится со скоростью геометрической прогрессии со знаменателем $q = 1/2$ и $c_0 = b - a$ (сравнить (2.43) с (2.61)).

Аналогично, метод простых итераций также является методом первого порядка сходимости и, следовательно, сходится со скоростью геометрической прогрессии со знаменателем $q = \max\limits_{x\in[a,b]} |\varphi'(x)|$ и $c_0 = |x_1 - x_0|/(1-q)$ (сравнить (2.59) с (2.61)).

В соответствии с леммой 2.3 и оценкой (2.49) итерационный метод Ньютона обладает вторым порядком сходимости и, следовательно, сходится со скоростью показательной функции, имеющей основание $q = c\,|\xi - x_0|$ и показатель степени 2^k:

$$|\xi - x_k| < c_0 \cdot q^{2^k},$$

где $c = M_2/2m_1$, $c_0 = 1/c$, т.е. значительно быстрее методов половинного деления и простых итераций.

Для ускорения медленно сходящихся методов существуют различные способы ускорения, один из которых, процесс Эйткена, основан на использовании нескольких предыдущих итерационных значений. Рассмотрим процесс ускорения по Эйткену применительно к методу простых итераций.

Из выражения (2.59) для оценки погрешности в методе простых итераций можно записать: $x_k - \xi = Aq^k$, где А — константа, не зависящая от q. Тогда это выражение для трех соседних итераций можно записать следующим образом:

$$\begin{cases} x_{k+1} - \xi = Aq^{k+1}, \\ x_k - \xi = Aq^k, \\ x_{k-1} - \xi = Aq^{k-1}, \end{cases}$$

откуда путем деления 1-го равенства на 2-е и 2-го на 3-е получим

$$\frac{x_{k+1} - \xi}{x_k - \xi} = \frac{x_k - \xi}{x_{k-1} - \xi}.$$

Решая это уравнение относительно ξ:

$$\xi = \frac{x_{k+1}x_{k-1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}},$$

получим почти точное значение корня уравнения (2.40). Это значение ξ можно принять за уточненное значение корня уравнения (2.40) по методу Эйткена на $(k+1)$-й итерации:

$$(x_{k+1})_{\text{ут}} = \frac{x_{k-1}x_{k+1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}}. \tag{2.62}$$

Можно показать, что погрешность метода простых итераций с использованием процедуры Эйткена теперь пропорциональна не q^k, как в (2.59), а q^{2^k}:

$$|\xi - x_k| \leqslant A q^{2^k},$$

т. е. процесс итераций сходится со скоростью показательной функции.

2.2.4. Замечания к методам отделения корней. Если уравнение $f(x) = 0$ является алгебраическим уравнением целой степени,

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0, \qquad (2.63)$$

то при отделении корней этого уравнения могут использоваться теоремы общей алгебры [1].

1. Основная теорема алгебры. *Число корней уравнения* (2.63) (*действительных, кратных, комплексных*) *в точности равно* n — *степени уравнения* (2.63), *причем если коэффициенты* $a_0, a_1, \ldots, a_n$ *все действительные, то возможные комплексные корни попарно сопряжены.*

2. Теорема Декарта. *Число положительных действительных корней уравнения* (2.63) *равно числу перемен знаков последовательности коэффициентов* $a_0, a_1, \ldots, a_n$ (*не считая нулевых коэффициентов*) *или меньше этого числа на четное число.*

Следствие. Число действительных отрицательных корней уравнения (2.63) равно числу постоянства знаков в последовательности коэффициентов $a_0, a_1, \ldots, a_n$, не считая нулевых, или меньше этого числа на четное число.

3. Теорема Гюа. *Если все корни уравнения* (2.63) *действительны, то в последовательности коэффициентов* $a_0, a_1, \ldots, a_n$ *квадраты некрайних коэффициентов больше произведения соседних, т. е.*

$$a_k^2 > a_{k-1} a_{k+1}, \quad k = 1, 2, \ldots, n-1. \qquad (2.64)$$

Следствие. Если хотя бы для одного некрайнего коэффициента условие (2.64) не выполняется, т. е. $a_k^2 \leqslant a_{k-1} a_{k+1}$, то имеется хотя бы одна пара комплексных корней.

Пример 2.14. Определить количество действительных и комплексных корней уравнения

$$x^4 + 8x^3 - 12x^2 + 104x - 20 = 0.$$

Решение.

Имеется три перемены знака. Значит, по теореме Декарта количество положительных корней должно быть равно одному или трем. Из 4-х корней по следствию из теоремы Гюа имеется, по крайней мере, одна пара комплексных корней, так как $(-12)^2 < 8 \cdot 104)$. Следовательно, среди действительных корней один положительный и один отрицательный.

§ 2.3. Численные методы решения систем нелинейных уравнений

При численном решении систем нелинейных уравнений предполагается существование искомого вектора решений, возможность нахождения начального приближения искомого вектора и, наконец, применение эффективных процедур уточнения решений. Пусть дана система нелинейных уравнений

$$\begin{cases} F_1(x_1, x_2, \dots, x_n) = 0, \\ F_2(x_1, x_2, \dots, x_n) = 0, \\ \dots\dots\dots\dots\dots\dots\dots \\ F_n(x_1, x_2, \dots, x_n) = 0, \end{cases} \tag{2.65}$$

или

$$F(x) = \vartheta, \tag{2.66}$$

где

$$F = \begin{pmatrix} F_1 & F_2 & \dots & F_n \end{pmatrix}^T,$$
$$x = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}^T,$$
$$\vartheta = \begin{pmatrix} 0 & 0 & \dots & 0 \end{pmatrix}^T,$$

и пусть имеется начальное приближение вектора неизвестных:

$$x^{(0)} = \begin{pmatrix} x_1^{(0)} \, x_2^{(0)} \dots x_n^{(0)} \end{pmatrix}^T. \tag{2.67}$$

Рис. 2.10. Начальное приближение в случае системы двух нелинейных уравнений

В случае системы двух нелинейных уравнений начальное приближение находят графически как точки пересечения графиков функций $F_1(x_1, x_2) = 0$ и $F_2(x_1, x_2) = 0$ (рис. 2.10).

2.3.1. Метод простых итераций и метод Зейделя решения систем нелинейных уравнений. Для использования метода простых итераций или метода Зейделя система (2.65) записывается в следующей *эквивалентной* форме:

$$x_i = \Phi_i(x_1, x_2, \ldots, x_n), \quad i = \overline{1, n}, \tag{2.68}$$

где $\Phi_i(x_1, x_2, \ldots, x_n)$ — *эквивалентные* функции.

Эквивалентные функции могут быть построены, например, в виде

$$\Phi_i(x_1, x_2, \ldots, x_n) = x_i + F_i(x_1, x_2, \ldots, x_n), \quad i = \overline{1, n}.$$

Тогда, если известно начальное приближение $x^0 = (x_1^{(0)} \ x_2^{(0)} \ \ldots \ x_n^{(0)})^T$, можно построить алгоритм метода простых итераций:

$$x_i^{(k+1)} = \Phi_i\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \quad i = \overline{1, n}, \tag{2.69}$$

или алгоритм метода Зейделя:

$$\begin{cases} x_1^{(k+1)} = \Phi_1\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \\ x_2^{(k+1)} = \Phi_2\left(x_1^{(k+1)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = \Phi_n\left(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, x_n^{(k)}\right). \end{cases} \tag{2.70}$$

Можно показать, что метод простых итераций (соответственно метод Зейделя) сходится к решению системы (2.65), если какая-либо норма матрицы Якоби, построенной по правым частям $\Phi_i\left(x^{(k)}\right)$, $i = \overline{1, n}$, эквивалентной системы (2.68), меньше единицы на каждой итерации:

$$\left\| J\left(x^{(k)}\right) \right\| < 1, \quad J\left(x^{(k)}\right) = \left[\frac{\partial \Phi_i\left(x^{(k)}\right)}{\partial x_j} \right]_{x = x^k}, \tag{2.71}$$
$$i, j = \overline{1, n}, \quad k = 0,\ 1,\ 2, \ldots$$

Метод простых итераций (2.69) (или Зейделя (2.70)) останавливается при выполнении условия

$$\left\| x^{(k+1)} - x^{(k)} \right\| \leqslant \varepsilon. \tag{2.72}$$

Тогда искомый вектор $x \approx x^{(k+1)}$.

Таким образом, методы простых итераций и Зейделя применяются для дифференцируемых функций $F_i(x)$, $i = \overline{1, n}$, в системе (2.65).

Пример 2.15. Методом простых итераций с точностью $\varepsilon = 10^{-2}$ найти решение нелинейной системы:

$$\begin{cases} \sin(x - 0{,}6) - y - 1{,}6 = 0, \\ 3x - \cos y - 0{,}9 = 0. \end{cases}$$

Соответствующая эквивалентная система имеет вид:

$$\begin{cases} x = 0{,}3 + 1/3 \cos y \equiv \Phi_1(x, y), \\ y = -1{,}6 + \sin(x - 0{,}6) \equiv \Phi_2(x, y). \end{cases}$$

Решение.

1) Начальное приближение $x^{(0)} = 0{,}15$, $y^{(0)} = -2{,}1$ определяется как точка пересечения графиков эквивалентных функций. $\left\| J\left(x^{(0)}, y^{(0)}\right) \right\|_1 = 0{,}31 < 1$.

2) $x^{(1)} = 0{,}1317$, $y^{(1)} = -2{,}035$.

Продолжая итерационный процесс, на 6-й и 7-й итерациях получим соответственно $x^{(6)} = 0{,}1539$; $y^{(6)} = -2{,}034$; $x^{(7)} = 0{,}151$; $y^{(7)} = -2{,}0319$; $\left\| \left(x^{(7)} y^{(7)}\right)^T - \left(x^{(6)} y^{(6)}\right)^T \right\| = 0{,}0029 < \varepsilon$.

УПРАЖНЕНИЯ.

2.41. Когда можно применять методы простых итераций и Зейделя?

Методом простых итераций решить следующие системы уравнений с точностью $\varepsilon = 10^{-2}$:

2.42. $\begin{cases} \cos(x - 1) + y = 0{,}5, \\ x - \cos y = 3; \end{cases}$

2.43. $\begin{cases} \sin x + 2y = 2, \\ \cos(y - 1) + x = 0{,}7; \end{cases}$

2.44. $\begin{cases} \cos x + y = 1{,}5, \\ 2x - \sin(y - 0{,}5) = 1; \end{cases}$

2.45. $\begin{cases} \cos(y - 2) + x = 0, \\ \sin(x + 0{,}5) - y = 1; \end{cases}$

2.46. $\begin{cases} 2y - \cos(x+1) = 0, \\ x + \sin y = -0{,}4; \end{cases}$

2.47. $\begin{cases} \cos(x+0{,}5) - y = 2, \\ \sin y - 2x = 1. \end{cases}$

2.3.2. Метод Ньютона. Рассмотрим систему нелинейных уравнений (2.65). Метод Ньютона решения систем вида (2.65) сводится к последовательному решению систем линейных алгебраических уравнений, полученных путем линеаризации системы нелинейных уравнений.

Для k-го приближения вектора неизвестных $x^{(k)} = \left(x_1^{(k)}\, x_2^{(k)}\, \ldots\, x_n^{(k)}\right)^T$ будем искать $(k+1)$-е приближение в виде

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k)}, \tag{2.73}$$

где вектор приращений $\Delta x^{(k)} = \left(\Delta x_1^{(k)}\, \Delta x_2^{(k)}\, \ldots\, \Delta x_n^{(k)}\right)^T$ подлежит определению.

Для нахождения этих приращений разложим функцию $F_1\left(x^{(k+1)}\right)$ в ряд Тейлора в окрестности точки $x^{(k)}$ до производных первого порядка включительно (оставим только линейную часть относительно Δx) и приравняем это разложение нулю: $F_1\left(x^{(k+1)}\right) = 0$, получим

$$F_1\left(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_n^{(k+1)}\right) = \\ = F_1\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right) + \frac{\partial F_1\left(x^{(k)}\right)}{\partial x_1}\Delta x_1^{(k)} + \\ + \frac{\partial F_1\left(x^{(k)}\right)}{\partial x_2}\Delta x_2^{(k)} + \ldots + \frac{\partial F_1\left(x^{(k)}\right)}{\partial x_n}\Delta x_n^{(k)} \approx 0.$$

Осуществляя те же процедуры для остальных уравнений системы (2.65), получим следующую СЛАУ относительно вектора неизвестных $\Delta x^{(k)} = \left(\Delta x_1^{(k)}\, \Delta x_2^{(k)}\, \ldots\, \Delta x_n^{(k)}\right)$:

$$
\begin{cases}
\dfrac{\partial F_1\left(x^{(k)}\right)}{\partial x_1}\Delta x_1^{(k)} + \dfrac{\partial F_1\left(x^{(k)}\right)}{\partial x_2}\Delta x_2^{(k)} + \ldots \\
\qquad\qquad \ldots + \dfrac{\partial F_1\left(x^{(k)}\right)}{\partial x_n}\Delta x_n^{(k)} = -F_1\left(x^{(k)}\right), \\
\dfrac{\partial F_2\left(x^{(k)}\right)}{\partial x_1}\Delta x_1^{(k)} + \dfrac{\partial F_2\left(x^{(k)}\right)}{\partial x_2}\Delta x_2^{(k)} + \ldots \\
\qquad\qquad \ldots + \dfrac{\partial F_2\left(x^{(k)}\right)}{\partial x_n}\Delta x_n^{(k)} = -F_2\left(x^{(k)}\right), \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\dfrac{\partial F_n\left(x^{(k)}\right)}{\partial x_1}\Delta x_1^{(k)} + \dfrac{\partial F_n\left(x^{(k)}\right)}{\partial x_2}\Delta x_2^{(k)} + \ldots \\
\qquad\qquad \ldots + \dfrac{\partial F_n\left(x^{(k)}\right)}{\partial x_n}\Delta x_n^{(k)} = -F_n\left(x^{(k)}\right).
\end{cases}
\tag{2.74}
$$

Для существования единственного решения СЛАУ (2.74) необходимо и достаточно, чтобы матрица этой СЛАУ

$$
J\left(x^{(k)}\right) = \begin{pmatrix} \dfrac{\partial F_1\left(x^{(k)}\right)}{\partial x_1} & \ldots & \dfrac{\partial F_1\left(x^{(k)}\right)}{\partial x_n} \\ \ldots & \ldots & \ldots \\ \dfrac{\partial F_n\left(x^{(k)}\right)}{\partial x_1} & \ldots & \dfrac{\partial F_n\left(x^{(k)}\right)}{\partial x_n} \end{pmatrix}
$$

была невырожденной $\left(\det J\left(x^{(k)}\right) \neq 0\right)$, т. е. существовала обратная матрица $\left(J\left(x^{(k)}\right)\right)^{-1}$. Матрицу J называют матрицей Якоби для системы (2.65), а ее определитель — якобианом этой системы.

Запишем (2.74) в векторно-матричной форме:

$$
J\left(x^{(k)}\right)\Delta x^{(k)} = -F\left(x^{(k)}\right),
$$

откуда

$$\Delta x^{(k)} = -J^{-1}\left(x^{(k)}\right) F\left(x^{(k)}\right). \qquad (2.75)$$

Подставляя (2.75) в (2.73), получим алгоритм Ньютона решения систем нелинейных уравнений (2.65):

$$x^{(k+1)} = x^{(k)} - J^{-1}\left(x^{(k)}\right) F\left(x^{(k)}\right) \qquad (2.76)$$

(сравнить с методом Ньютона для одного нелинейного уравнения $x^{(k+1)} = x^{(k)} - f\left(x^{(k)}\right)/f'\left(x^{(k)}\right)$).

Итерационный процесс (2.76) заканчивается при выполнении условия

$$\left\|x^{(k+1)} - x^{(k)}\right\| \leqslant \varepsilon,$$

где ε — заданная точность, а вектор $x \approx x^{(k+1)}$.

Таким образом, для применения метода Ньютона (2.76) необходимо выполнение следующих двух условий:

— существование частных производных первого порядка от функций $F_i(x_1, x_2, \ldots, x_n)$, $i = \overline{1, n}$, по всем переменным x_j, $j = \overline{1, n}$;

— матрица Якоби для системы (2.65) на каждой итерации $k = 0, 1, 2, \ldots$ должна быть невырождена.

Пример 2.16. Методом Ньютона с точностью $\varepsilon = 10^{-2}$ решить систему нелинейных уравнений

$$\begin{cases} F_1(x_1, x_2) \equiv x_1 - 2x_2^2 + 1 = 0, \\ F_2(x_1, x_2) \equiv -x_1^2 + 2x_2 - 1 = 0. \end{cases}$$

Решение.

Графически можно показать, что первое вектор-решение расположено в прямоугольнике $-0{,}8 < (x_1)_{\text{I}} < 0{,}4$; $0{,}5 < (x_2)_{\text{I}} < 1{,}0$, а второе — в прямоугольнике $0{,}8 < (x_1)_{\text{II}} < 1{,}1$; $0{,}8 < (x_2)_{\text{II}} < 1{,}1$. Уточним второе вектор-решение.

1) Начальное приближение $x_1^{(0)} = (0{,}8 + 1{,}1)/2 = 0{,}95$; $x_2^{(0)} = (0{,}8 + 1{,}1)/2 = 0{,}95$;

2) $J\left(x^{(0)}\right) = \begin{bmatrix} 1 & -3{,}8 \\ -1{,}9 & 2{,}0 \end{bmatrix}$; $J^{-1}\left(x^{(0)}\right) = \dfrac{1}{-5{,}22}\begin{bmatrix} 2{,}0 & 3{,}8 \\ 1{,}9 & 1{,}0 \end{bmatrix}$;

3) $\left(x_1^{(1)} x_2^{(1)}\right)^T = \left(x_1^{(0)} x_2^{(0)}\right)^T - J^{-1}\left(x^{(0)}\right)\left(F_1\left(x^{(0)}\right) F_2\left(x^{(0)}\right)\right)^T =$
$= (1{,}00374\ \ 1{,}0023)^T$;

4) $\left(x_1^{(2)}\ x_2^{(2)}\right)^T = \left(x_1^{(1)}\ x_2^{(1)}\right)^T - J^{-1}\left(x^{(1)}\right) F\left(x^{(1)}\right) =$
$= (1{,}0013\ 1{,}0034)$; $\left\|x^{(2)} - x^{(1)}\right\| = 0{,}0024 < \varepsilon$.

УПРАЖНЕНИЯ.

2.48. В каких случаях нельзя применять метод Ньютона?

Методом Ньютона с точностью $\varepsilon = 10^{-3}$ найти векторы-решения для следующих нелинейных систем [6]:

2.49. $\begin{cases} \sin(x+1) - y = 1, \\ 2x + \cos y = 2; \end{cases}$

2.50. $\begin{cases} \operatorname{tg}(xy + 0{,}2) = x^2, \\ x^2 + 2y^2 = 1; \end{cases}$

2.51. $\begin{cases} 2x + \operatorname{tg}(xy) = 0, \\ \left(y^2 - 6\right)^2 + \ln x = 0; \end{cases}$

2.52. $\begin{cases} 0{,}6x + 7{,}5y + x^2 y = 0, \\ \cos y + 6x = 0; \end{cases}$

2.53. $\begin{cases} \sin(x + 0{,}8) + 2y - 1 = 0, \\ \cos(y + 0{,}6) + 0{,}6x = 0; \end{cases}$

2.54. $\begin{cases} \operatorname{tg} x - \cos(1{,}5y) = 0, \\ 2y^3 - x^2 + 4x - 3 = 0. \end{cases}$

2.55. Сколько итераций требуется в методе Ньютона для СЛАУ?

§ 2.4. Численные методы решения задач на собственные значения и собственные векторы матриц линейных преобразований

2.4.1. Основные определения и спектральные свойства матриц. Рассмотрим матрицу $A_{n\times n}$ в n-мерном вещественном пространстве R^n векторов

$$x = (x_1\ x_2\ \ldots\ x_n)^T.$$

1. Собственным вектором x матрицы A называется *ненулевой* вектор ($x \neq \vartheta$), удовлетворяющий равенству

$$Ax = \lambda x, \tag{2.77}$$

где λ — собственное значение матрицы A, соответствующее рассматриваемому собственному вектору.

2. Собственные значения матрицы A с действительными элементами могут быть вещественными различными, вещественными кратными, комплексными попарно сопряженными, комплексными кратными.

3. Классический способ нахождения собственных значений и собственных векторов известен и заключается в следующем:
— для однородной СЛАУ, полученной из (2.77),

$$(A - \lambda E)\, x = \vartheta, \quad \vartheta = (0\ 0\ \ldots\ 0)^T, \tag{2.78}$$

ненулевые решения ($x \neq \vartheta$, а именно такие решения и находятся) имеют место при

$$\det (A - \lambda E) = 0, \tag{2.79}$$

причем уравнение (2.79) называют *характеристическим уравнением*, а выражение в левой части — *характеристическим многочленом*;
— каким-либо способом находят решения $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_n$ алгебраического уравнения (2.79) n-й степени (предположим, что они вещественны и различны);
— решая однородную СЛАУ (2.78) для различных собственных значений $\lambda_j,\ j = \overline{1, n}$,

$$(A - \lambda_j E)\, x^j = \vartheta, \quad j = \overline{1, n},$$

получаем линейно независимые собственные векторы x^j, $j = \overline{1, n}$, соответствующие собственным значениям λ_j, $j = \overline{1, n}$.

4. Если количество линейно независимых собственных векторов матрицы $A_{n \times n}$ совпадает с размерностью пространства R^n, то их можно принять за новый базис, в котором матрица $A_{n \times n}$ примет диагональный вид

$$\Lambda = U^{-1} \cdot A \cdot U. \tag{2.80}$$

На главной диагонали матрицы Λ находятся собственные значения, а столбцы матрицы преобразования U являются собственными векторами матрицы A (матрицы Λ и A, удовлетворяющие равенству (2.80), называются *подобными*).

5. ***Основные свойства*** собственных значений и собственных векторов:

а) ***Теорема Перрона.*** *Если все элементы квадратной матрицы положительны, то наибольшее по модулю собственное значение ее также положительно, является простым (не кратным) корнем характеристического уравнения и ему соответствует собственный вектор с положительными компонентами*;

б) если для собственного значения λ_j найден собственный вектор x^j, то вектор cx^j, где c — произвольное число, также является собственным вектором, соответствующим собственному значению λ_j, при этом векторы x^j и cx^j являются линейно зависимыми.

в) попарно различным собственным значениям соответствуют линейно независимые собственные векторы;

г) k-кратному корню характеристического уравнения (2.79), построенного для *произвольной* матрицы $A_{n \times n}$, *соответствует не более* k ($\leqslant k$) линейно независимых собственных векторов;

д) собственные значения *подобных матриц* Λ и A, удовлетворяющих равенству (2.80), совпадают;

е) *симметрическая матрица* A $\left(A = A^T\right)$ имеет полный спектр λ_j , $j = \overline{1, n}$, *вещественных* собственных значений;

ж) *положительно определенная симметрическая матрица* $\left(A = A^T,\ (Ax, x) > 0\right)$ имеет *полный спектр вещественных положительных* собственных значений;

з) k-кратному корню характеристического уравнения (2.79) *симметрической* матрицы соответствует *ровно* k линейно независимых собственных векторов;

и) *симметрическая матрица* имеет ровно n ортогональных собственных векторов (в соответствии со свойством е), приняв которые за новый базис (т. е. построив матрицу преобразования U, взяв в качестве ее столбцов координатные столбцы собственных векторов), *можно преобразовать симметрическую матрицу A к диагональному виду* с помощью преобразования (2.80).

к) для *симметрической матрицы A* матрица преобразования U в (2.80) является ортогональной $U^{-1} = U^T$, и, следовательно, преобразование (2.80) имеет вид

$$\Lambda = U^T \cdot A \cdot U. \tag{2.81}$$

Покажем, что собственные значения матриц Λ и A в соотношении (2.80) (или (2.81)) одинаковы, собственные векторы — различны, а столбцы матрицы U — суть координатные столбцы собственных векторов матрицы A. Действительно, в соответствии с (2.80) характеристические многочлены матриц Λ и A одинаковы, поскольку

$$\det \Lambda = \frac{1}{\det U} \cdot \det A \cdot \det U = \det A$$

на основе свойства определителя произведения матриц и определителя обратной матрицы, т. е. собственные значения матриц Λ и A одинаковы.

Обозначим собственные векторы матрицы Λ через y:

$$\Lambda y = \lambda y, \tag{2.82}$$

а собственные векторы матрицы A — через x, тогда

$$\Lambda y = U^{-1} A U y = \lambda \cdot y, \quad A(Uy) = \lambda \cdot (Uy), \text{ но } Ax = \lambda x,$$

откуда

$$Uy = x. \tag{2.83}$$

Для первого собственного вектора $y^{(1)}$ матрицы Λ равенство (2.82) имеет вид

$$\begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} \begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ \vdots \\ y_n^{(1)} \end{pmatrix} = \lambda_1 \begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ \vdots \\ y_n^{(1)} \end{pmatrix},$$

$$\begin{cases} \lambda_1 y_1^{(1)} = \lambda_1 y_1^{(1)}, \quad \text{откуда} \;\; y_1^{(1)} = 1, \\ \lambda_2 y_2^{(1)} = \lambda_1 y_2^{(1)}, \quad \lambda_2 \neq \lambda_1, \quad \text{следовательно} \;\; y_2^{(1)} = 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \lambda_n y_n^{(1)} = \lambda_1 y_n^{(1)}, \quad \lambda_n \neq \lambda_1, \quad \text{следовательно} \;\; y_n^{(1)} = 0. \end{cases}$$

Таким образом,

$$y^{(1)} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}; \quad y^{(2)} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}; \quad \dots; \quad y^{(n)} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

Из (2.83) для $y^{(1)}$ имеем $Uy^{(1)} = x^{(1)}$, откуда

$$\begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ u_{n1} & u_{n2} & \dots & u_{nn} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_n^{(1)} \end{pmatrix}.$$

Аналогично для всех остальных векторов $y^{(j)}$, $j = \overline{2, n}$ получаем

$$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_n^{(1)} \end{pmatrix} = \begin{pmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{pmatrix} \dots \begin{pmatrix} x_1^{(n)} \\ x_2^{(n)} \\ \vdots \\ x_n^{(n)} \end{pmatrix} = \begin{pmatrix} u_{1n} \\ u_{2n} \\ \vdots \\ u_{nn} \end{pmatrix},$$

что и требовалось доказать, т. е. столбцы матрицы U суть собственные векторы матрицы A.

УПРАЖНЕНИЯ.

2.56. Показать, что для любого собственного значения $\lambda(A)$ *невырожденной* матрицы A справедлива оценка $1/\|A^{-1}\| \leqslant |\lambda(A)|\,\|A^{-1}\| \leqslant |\lambda(A)| \leqslant \|A\|$ (использовать определение $\mathrm{cond}(A)$ в п. 2.1.5).

2.57. Доказать, что для симметрической матрицы A столбцы матрицы U в равенстве (2.80) являются собственными ортогональными векторами (столбцы матрицы U ортогональны, если U — ортогональная матрица, т. е. $U^{-1} = U^T$).

2.58. Доказать, что собственные значения симметрической положительно определенной матрицы A $\left(A = A^T,\ (Ax,\, x) > 0\right)$ являются положительными числами.

2.59. Доказать, что у матрицы

$$A = \begin{bmatrix} 2 & 0{,}4 & 0{,}4 \\ 0{,}3 & 4 & 0{,}4 \\ 0{,}1 & 0{,}1 & 5 \end{bmatrix}$$

все собственные значения вещественны (использовать норму обеих частей равенства $Ax = \lambda x$). Указать интервалы, которым принадлежат собственные значения.

2.61. Показать, что для экстремальных собственных значений $\lambda_{\max}$, $\lambda_{\min}$ симметрической матрицы A справедливы оценки

$$\lambda_{\max}(A) \geqslant \max(a_{ii});$$

$$\lambda_{\min}(A) \leqslant \min(a_{ii}).$$

2.62. Показать, что у вещественной трехдиагональной матрицы A все собственные значения вещественны, если $a_{i+1}c_i > 0$, $i = \overline{1,\, n-1}$ (показать, что она подобна трехдиагональной симметрической матрице). Элементы i-й строки матрицы A — a_i, b_i, c_i, $i = \overline{1, n}$.

2.4.2. Метод вращений Якоби численного решения задач на собственные значения и собственные векторы матриц. Метод вращений Якоби применим только для симметрических матриц $A_{n\times n}$ $(A = A^T)$ и решает полную проблему собственных значений и собственных векторов таких матриц. Он основан на отыскании с помощью итерационных процедур матрицы U в преобразовании подобия [5,7] $\Lambda = U^{-1}AU$, а поскольку

для симметрических матриц A матрица преобразования подобия U является ортогональной ($U^{-1} = U^T$), то

$$\Lambda = U^T A U, \tag{2.84}$$

где Λ — диагональная матрица с собственными значениями на главной диагонали

$$\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix}.$$

Геометрически метод вращений представляет собой последовательное вращение системы базисных векторов в n-мерном вещественном пространстве R^n в плоскости некоторых, специальным образом выбранных двух векторов. Такая последовательность вращений осуществляется до тех пор, пока новые базисные векторы не совпадут с собственными векторами, т. е. в качестве базисных принимаются собственные векторы матрицы A. Поскольку симметрическая матрица имеет полный набор ортогональных собственных векторов (см. п. 2.4.1.), то для нее можно построить алгоритм метода вращений на основе преобразования подобия (2.84).

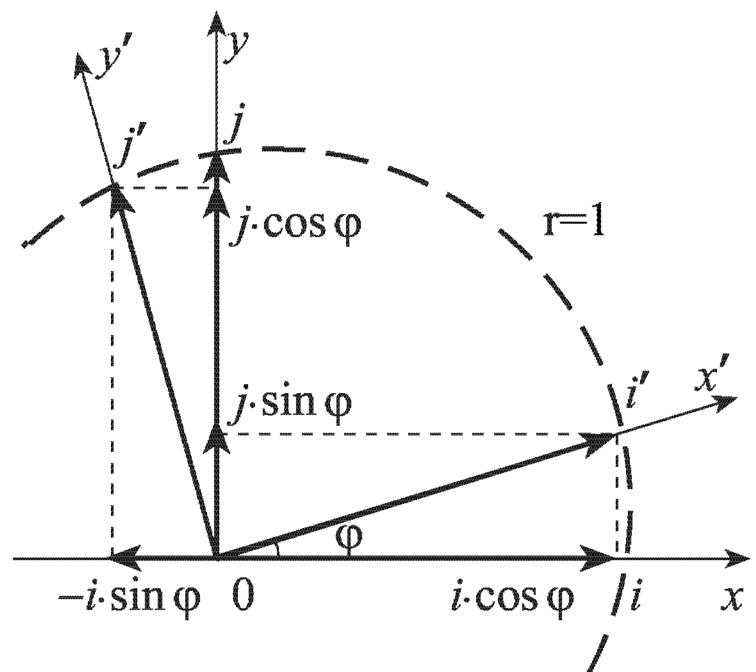


Рис. 2.11. Геометрическая интерпретация метода вращения на плоскости

Пример 2.17. На плоскости R^2 с ортонормированным базисом (i, j) построить матрицу преобразования вращения U около начала отсчета на угол φ.

Р е ш е н и е. В соответствии с рис. 2.11 имеем

$$\begin{cases} i' = i\cos\varphi + j\sin\varphi, \\ j' = -i\sin\varphi + j\cos\varphi, \end{cases}$$

$$(i'\,j') = (i\,j)\cdot\begin{pmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{pmatrix},$$

т. е. матрица вращения U на плоскости имеет вид

$$U = \begin{pmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{pmatrix}.$$

Пример 2.18. В пространстве R^3 с ортонормированным базисом $(i,\ j,\ k)$ построить матрицы вращения на угол φ соответственно в плоскостях $(j,\ k)$, $(k,\ i)$, $(i,\ j)$.

Решение. В [8] доказывается, что матрицы вращения U_x, U_y, U_z имеют вид

$$U_x = \begin{pmatrix}1 & 0 & 0\\ 0 & \cos\varphi & \boxed{-\sin\varphi}\\ 0 & \sin\varphi & \cos\varphi\end{pmatrix},\quad U_y = \begin{pmatrix}\cos\varphi & 0 & \sin\varphi\\ 0 & 1 & 0\\ \boxed{-\sin\varphi} & 0 & \cos\varphi\end{pmatrix},$$

$$U_z = \begin{pmatrix}\cos\varphi & \boxed{-\sin\varphi} & 0\\ \sin\varphi & \cos\varphi & 0\\ 0 & 0 & 1\end{pmatrix},$$

из которого видно, что номера базисных векторов в их упорядоченной совокупности $(i,\ j,\ k)$, находящихся в плоскости вращения, совпадают с номером строки и номером столбца, на пересечении которых стоит элемент $-\sin\varphi$.

Матрица полного вращения U около начала координат на некоторый угол определяется как преобразование

$$U = U_x U_y U_z.$$

Можно показать, что в n-мерном пространстве R^n в плоскости вращения, содержащей базисные векторы с номерами i и j, матрица вращения имеет вид

$$
U = \begin{pmatrix}
1 & & & \vdots & & & \vdots & & & \\
& \ddots & & \vdots & & 0 & \vdots & & & \\
& & 1 & \vdots & & & \vdots & & & \\
\ldots & \ldots & \ldots & \cos\varphi & \ldots & \ldots & \boxed{-\sin\varphi} & \ldots & \ldots & \\
& & & \vdots & 1 & & \vdots & & & \\
& 0 & & \vdots & & \ddots & \vdots & & 0 & \\
& & & \vdots & & 1 & \vdots & & & \\
\ldots & \ldots & \ldots & \sin\varphi & \ldots & \ldots & \cos\varphi & \ldots & \ldots & \\
& & & \vdots & & & \vdots & 1 & & \\
& 0 & & \vdots & & 0 & \vdots & & \ddots & \\
& & & \vdots & & & \vdots & & & 1
\end{pmatrix}
\begin{matrix} \\ \\ \\ i \\ \\ \\ \\ j \\ \\ \\ \\ \end{matrix}
$$

(столбцы i и j отмечены над матрицей)

Пусть дана симметрическая матрица A. Требуется вычислить для нее с точностью ε все собственные значения и соответствующие им собственные векторы. Алгоритм метода вращений следующий [5].

Пусть известна матрица $A^{(k)}$, являющаяся приближением матрицы Λ на k-й итерации, при этом для $k = 0$ $A^{(0)} = A$.

1. Выбирается максимальный по модулю недиагональный элемент $a_{ij}^{(k)}$ матрицы $A^{(k)}$ $\left(\max\limits_{i<j}\left|a_{ij}^{(k)}\right|\right)$.

2. Ставится задача найти такую ортогональную матрицу $U^{(k)}$, чтобы в результате преобразования подобия $A^{(k+1)} = U^{(k)T}A^{(k)}U^{(k)}$ произошло обнуление элемента $a_{ij}^{(k+1)}$ матрицы $A^{(k+1)}$. В качестве ортогональной матрицы выбирается представленная ниже матрица вращения $U^{(k)}$. В ней на пересечении i-й строки и j-го столбца находится элемент $u_{ij}^{(k)} = -\sin\varphi^{(k)}$, где $\varphi^{(k)}$ — угол вращения, подлежащий определению. Симметрично относительно главной диагонали (j-я строка, i-й столбец) расположен элемент $u_{ji}^{(k)} = \sin\varphi^{(k)}$. Диагональные элементы $u_{ii}^{(k)}$ и $u_{jj}^{(k)}$ равны соответственно $u_{ii}^{(k)} = \cos\varphi^{(k)}$, $u_{jj}^{k} = \cos\varphi^{(k)}$; другие диагональные элементы $u_{mm}^{(k)} = 1$, $m = \overline{1,n}$, $m \neq i$, $m \neq j$; остальные элементы в матрице вращения $U^{(k)}$ равны нулю.

Угол вращения $\varphi^{(k)}$ определяется из условия $a_{ij}^{(k+1)} = 0$:

$$\varphi^{(k)} = \frac{1}{2} \operatorname{arctg} \frac{2a_{ij}^{(k)}}{a_{ii}^{(k)} - a_{jj}^{(k)}}, \tag{2.85}$$

$$U^{(k)} = \begin{pmatrix} 1 & & & \vdots & & & \vdots & & & \\ & \ddots & & \vdots & & 0 & \vdots & & & \\ & & 1 & \vdots & & & \vdots & & & \\ \dots & \dots & \dots & \cos\varphi^{(k)} & \dots & \dots & -\sin\varphi^{(k)} & \dots & \dots & \dots \\ & & & \vdots & 1 & & \vdots & & & \\ & 0 & & \vdots & & \ddots & \vdots & & 0 & \\ & & & \vdots & & 1 & \vdots & & & \\ \dots & \dots & \dots & \sin\varphi^{(k)} & \dots & \dots & \cos\varphi^{(k)} & \dots & \dots & \dots \\ & & & \vdots & & & \vdots & 1 & & \\ & 0 & & \vdots & & 0 & \vdots & & \ddots & \\ & & & \vdots & & & \vdots & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ i \\ \\ \\ \\ j \\ \\ \\ \\ \end{matrix},$$

(столбцы i и j)

причем если $a_{ii}^{(k)} = a_{jj}^{(k)}$, то $\varphi^{(k)} = \dfrac{\pi}{4}$.

3. Строится матрица $A^{(k+1)}$:

$$A^{(k+1)} = U^{(k)\,T} A^{(k)} U^{(k)},$$

в которой элемент $a_{ij}^{(k+1)} \approx 0$.

В качестве критерия окончания итерационного процесса используется условие малости суммы квадратов внедиагональных элементов:

$$t\left(A^{(k+1)}\right) = \left(\sum_{i,j;\,i<j} \left(a_{ij}^{(k+1)}\right)^2\right)^{1/2}.$$

Если $t\left(A^{k+1}\right) > \varepsilon$, то итерационный процесс

$$A^{(k+1)} = U^{(k)\,T} A^{(k)} U^{(k)} = U^{(k)\,T} U^{(k-1)\,T} \dots$$
$$\dots U^{(0)\,T} A^{(0)} U^{(0)} U^{(1)} \dots U^{(k)}$$

продолжается. Если $t\left(A^{k+1}\right) < \varepsilon$, то итерационный процесс останавливается, и в качестве искомых собственных значений принимаются: $\lambda_1 \approx a_{11}^{(k+1)}$, $\lambda_2 \approx a_{22}^{(k+1)}$, ..., $\lambda_n \approx a_{nn}^{(k+1)}$.

Координатными столбцами собственных векторов матрицы A в единичном базисе будут столбцы матрицы $U = U^{(0)}U^{(1)}\ldots U^{(k)}$, т. е.

$$\left(x^1\right)^{\mathrm{T}} = \left(u_{11}\,u_{21}\ \ldots\ u_{n1}\right), \quad \left(x^2\right)^{\mathrm{T}} = \left(u_{12}\,u_{22}\ \ldots\ u_{n2}\right), \quad \ldots$$
$$\ldots, \quad \left(x^n\right)^{\mathrm{T}} = \left(u_{1n}\,u_{2n}\ \ldots\ u_{nn}\right),$$

причем эти собственные векторы будут ортогональны между собой, т. е. $\left(x^l,\, x^m\right) \approx 0$, $l \neq m$.

Остановимся подробнее на вычислении угла вращения.

Для определения угла вращения $\varphi^{(k)}$ на k-ой итерации приравняем нулю внедиагональный элемент $a_{ij}^{(k+1)}$ (стоящий на месте максимального по модулю элемента $a_{ij}^{(k)}$ в матрице $A^{(k)}$) в произведении $A^{(k+1)} = U^{(k)T}A^{(k)}U^{(k)}$, получим

$$A^{(k+1)} = U^{(k)T}A^{(k)}U^{(k)} = \begin{pmatrix} 1 & & & \vdots & & & \vdots & & & \\ & \ddots & & \vdots & & 0 & \vdots & & & \\ & & 1 & \vdots & & & \vdots & & & \\ \ldots & \ldots & \ldots & \cos\varphi^{(k)} & \ldots & \ldots & \sin\varphi^{(k)} & \ldots & \ldots & \ldots \\ & & & \vdots & 1 & & \vdots & & & \\ & 0 & & \vdots & & \ddots & \vdots & & 0 & \\ & & & \vdots & & 1 & \vdots & & & \\ \ldots & \ldots & \ldots & -\sin\varphi^{(k)} & \ldots & \ldots & \cos\varphi^{(k)} & \ldots & \ldots & \ldots \\ & & & \vdots & & & \vdots & 1 & & \\ & 0 & & \vdots & & 0 & \vdots & & \ddots & \\ & & & \vdots & & & \vdots & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ i \\ \\ \times \\ \\ j \\ \\ \\ \\ \end{matrix}$$

(столбцы i и j отмечены над матрицей)

$$\times \begin{pmatrix} a_{11}^{(k)} \cdots a_{1i}^{(k)} \cdots a_{1j}^{(k)} \cdots a_{1n}^{(k)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{i1}^{(k)} \cdots a_{ii}^{(k)} \cdots a_{ij}^{(k)} \cdots a_{in}^{(k)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{j1}^{(k)} \cdots a_{ji}^{(k)} \cdots a_{jj}^{(k)} \cdots a_{jn}^{(k)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}^{(k)} \cdots a_{ni}^{(k)} \cdots a_{nj}^{(k)} \cdots a_{nn}^{(k)} \end{pmatrix} \times$$

$$\times \begin{pmatrix} 1 & & \vdots & & \vdots & & \\ & \ddots & \vdots & 0 & \vdots & & \\ & & 1 \vdots & & \vdots & & \\ \ldots & \ldots & \cos\varphi^{(k)} & \ldots & -\sin\varphi^{(k)} & \ldots & \ldots \\ & & \vdots & 1 & \vdots & & \\ & 0 & \vdots & \ddots & \vdots & 0 & \\ & & \vdots & 1 & \vdots & & \\ \ldots & \ldots & \sin\varphi^{(k)} & \ldots & \cos\varphi^{(k)} & \ldots & \ldots \\ & & \vdots & & \vdots & 1 & \\ & 0 & \vdots & 0 & \vdots & \ddots & \\ & & \vdots & & \vdots & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ i \\ \\ \\ \\ j \\ \\ \\ \\ \end{matrix} =$$

$$= \begin{pmatrix} \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots \\ \vdots & \vdots\, a_{ij}^{(k+1)} & \vdots \\ \ldots & \ldots & \ldots \\ \vdots\, a_{ji}^{(k+1)} & \vdots & \vdots \\ \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots \end{pmatrix} \begin{matrix} \\ \\ i \\ \\ j \\ \\ \\ \end{matrix}$$

(столбцы i и j отмечены над матрицами.)

После перемножения на месте элемента a_{ij}^{k+1} будет следующее выражение, которое приравняется нулю:

$$a_{ij}^{(k+1)} = (-a_{ii}^{(k)} \sin\varphi^{(k)} + a_{ij}^{(k)} \cos\varphi^{(k)}) \cos\varphi^{(k)} + \\ + (-a_{ji}^{(k)} \sin\varphi^{(k)} + a_{jj}^{(k)} \cos\varphi^{(k)}) \sin\varphi^{(k)} = 0,$$

а поскольку $A^{(k)}$ — симметрическая матрица (т. е. $a_{ij}^{(k)} = a_{ji}^{(k)}$), из последнего равенства следует выражение (2.85).

Пример 2.19. С точностью $\varepsilon = 0,3$ вычислить собственные значения и собственные векторы матрицы

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & \mathbf{3} \\ 1 & 3 & 6 \end{bmatrix} \equiv A^{(0)}.$$

Р е ш е н и е.

1) Выбираем максимальный по модулю внедиагональный элемент матрицы $A^{(0)}$, т. е. находим $a_{ij}^{(0)}$ такой, что $\left|a_{ij}^{(0)}\right| = \max\limits_{i<j} \left|a_{ij}^{(0)}\right|$. Им является элемент $a_{23}^{(0)} = 3$, т. е. $i = 2$, $j = 3$.

2) Находим соответствующую этому элементу матрицу вращения:

$$\begin{array}{c} \qquad\qquad\qquad j=3 \\ U^{(0)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi^{(0)} & -\sin\varphi^{(0)} \\ 0 & \sin\varphi^{(0)} & \cos\varphi^{(0)} \end{bmatrix} i = 2; \end{array}$$

$$\varphi^{(0)} = \frac{1}{2}\operatorname{arctg}\frac{2\cdot 3}{5-6} = -0{,}7033;\ \sin\varphi^{(0)} = -0{,}65;\ \cos\varphi^{(0)} = 0{,}76;$$

$$U^{(0)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0{,}76 & 0{,}65 \\ 0 & -0{,}65 & 0{,}76 \end{bmatrix}.$$

3) Вычисляем матрицу $A^{(1)}$:

$$A^{(1)} = U^{(0)\,T} A^{(0)} U^{(0)} = \begin{bmatrix} 4 & 0{,}87 & \mathbf{2{,}06} \\ 0{,}87 & 2{,}46 & -0{,}03 \\ 2{,}06 & -0{,}03 & 8{,}54 \end{bmatrix}.$$

В полученной матрице с точностью до ошибок округления элемент $a_{23}^{(1)} \approx 0$. Проверяется окончание итерационного процесса:

$$t\left(A^{(1)}\right) = \left(\sum_{i,j;\,i<j}\left(a_{ij}^{(1)}\right)^2\right)^{1/2} = \left(0{,}87^2 + 2{,}06^2 + (-0{,}03)^2\right)^{1/2} > \varepsilon,$$

следовательно итерационный процесс необходимо продолжить.

Переходим к следующей итерации ($k = 1$):

$$a_{13}^{(1)} = 2{,}06; \quad \left(\left|a_{13}^{(1)}\right| = \max_{i,j;\ i<j} \left|a_{ij}^{(1)}\right|\right), \ \text{т. е.}\ \ i = 1,\ j = 3;$$

$$U^{(1)} = \begin{bmatrix} \cos\varphi^{(1)} & 0 & -\sin\varphi^{(1)} \\ 0 & 1 & 0 \\ \sin\varphi^{(1)} & 0 & \cos\varphi^{(1)} \end{bmatrix} \begin{matrix} i = 1 \\ \\ \\ \end{matrix} ; \quad (j = 3)$$

$$\varphi^{(1)} = \frac{1}{2}\operatorname{arctg}\frac{2 \cdot 2{,}06}{4 - 8{,}54} = -0{,}3693;$$

$$\sin\varphi^{(1)} = -0{,}361; \quad \cos\varphi^{(1)} = 0{,}933;$$

$$U^{(1)} = \begin{bmatrix} 0{,}933 & 0 & 0{,}361 \\ 0 & 1 & 0 \\ -0{,}361 & 0 & 0{,}933 \end{bmatrix};$$

$$A^{(2)} = U^{(1)\ T} A^{(1)} U^{(1)} = \begin{bmatrix} 3{,}19 & \mathbf{0{,}819} & 0{,}005 \\ 0{,}819 & 2{,}46 & 0{,}28 \\ 0{,}005 & 0{,}28 & 9{,}38 \end{bmatrix}.$$

$$t\left(A^{(2)}\right) = \left(\sum_{i,j;\, i<j} \left(a_{ij}^{(2)}\right)^2\right)^{1/2} = \left(0{,}819^2 + 0{,}28^2 + 0{,}005^2\right)^{1/2} > \varepsilon.$$

Переходим к следующей итерации ($k = 2$):

$$a_{12}^{(2)} = 0{,}819; \quad \left(\left|a_{12}^{(2)}\right| = \max_{i,j;\ i<j} \left|a_{ij}^{(2)}\right|\right), \ \text{т. е.}\ \ i = 1,\ j = 2;$$

$$\varphi^{(2)} = \frac{1}{2}\,\mathrm{arctg}\,\frac{2 \cdot 0{,}819}{3{,}19 - 2{,}46} = 0{,}5758;$$

$$\sin\varphi^{(2)} = 0{,}5445; \quad \cos\varphi^{(2)} = 0{,}8388.$$

$$U^{(2)} = \begin{bmatrix} 0{,}8388 & -0{,}5445 & 0 \\ 0{,}5445 & 0{,}8388 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{matrix} i = 1 \\ \\ \\ \end{matrix} ; \qquad (j = 2)$$

$$A^{(3)} = U^{(2)\,T} A^{(2)} U^{(2)} = \begin{bmatrix} 3{,}706 & 0{,}0003 & 0{,}1565 \\ 0{,}0003 & 1{,}929 & 0{,}232 \\ 0{,}1565 & 0{,}232 & 9{,}38 \end{bmatrix};$$

$$t\left(A^{(3)}\right) = (0{,}0003^2 + 0{,}1565^2 + 0{,}232^2)^{1/2} = 0{,}07839^{1/2} < \varepsilon.$$

Таким образом, в качестве искомых собственных значений принимаются диагональные элементы матрицы $A^{(3)}$:

$$\lambda_1 \approx 3{,}706; \quad \lambda_2 \approx 1{,}929; \quad \lambda_3 \approx 9{,}38.$$

Собственные векторы определяются из произведения

$$U^{(0)} U^{(1)} U^{(2)} = \begin{bmatrix} 0{,}78 & -0{,}5064 & 0{,}361 \\ 0{,}2209 & 0{,}7625 & 0{,}6 \\ -0{,}58 & -0{,}398 & 0{,}7 \end{bmatrix};$$

$$x^1 = \begin{bmatrix} 0{,}78 \\ 0{,}2209 \\ -0{,}58 \end{bmatrix}; \quad x^2 = \begin{bmatrix} -0{,}5064 \\ 0{,}7625 \\ -0{,}398 \end{bmatrix}; \quad x^3 = \begin{bmatrix} 0{,}361 \\ 0{,}6 \\ 0{,}7 \end{bmatrix}.$$

Полученные собственные векторы ортогональны в пределах заданной точности, т. е.

$$\left(x^1,\, x^2\right) = -0{,}00384; \quad \left(x^1, x^3\right) = 0{,}0081; \quad \left(x^2,\, x^3\right) = -0{,}0039.$$

УПРАЖНЕНИЯ.

2.63. Показать, что собственные значения матриц Λ и A в преобразовании подобия (2.84) одинаковы, а собственные векторы — нет.

2.64. Показать, что для симметрической матрицы A матрица преобразования подобия U в (2.84) ортогональна $\left(U^{-1} = U^T\right)$.

2.65. Для следующих матриц с точностью $\varepsilon = 10^{-2}$ определить методом вращения собственные значения и собственные векторы:

$$\text{а)} \begin{bmatrix} 1{,}0 & 1{,}3 & 1{,}2 \\ 1{,}3 & 0{,}6 & 1{,}5 \\ 1{,}2 & 1{,}5 & 0{,}8 \end{bmatrix}; \qquad \text{б)} \begin{bmatrix} 0{,}6 & 1{,}0 & 2{,}1 \\ 1{,}0 & 1{,}4 & 0{,}5 \\ 2{,}1 & 0{,}5 & 1{,}0 \end{bmatrix};$$

$$\text{в)} \begin{bmatrix} 2{,}2 & 0{,}3 & 0{,}5 \\ 0{,}3 & 1{,}2 & 0{,}6 \\ 0{,}5 & 0{,}6 & 1{,}0 \end{bmatrix}; \qquad \text{г)} \begin{bmatrix} 1{,}2 & 1{,}5 & 0{,}3 \\ 1{,}5 & 0{,}4 & 2{,}0 \\ 0{,}3 & 2{,}0 & 1{,}2 \end{bmatrix};$$

$$\text{д)} \begin{bmatrix} 1{,}3 & 0{,}6 & 0{,}8 \\ 0{,}6 & 1{,}0 & 1{,}2 \\ 0{,}8 & 1{,}2 & 1{,}5 \end{bmatrix}; \qquad \text{е)} \begin{bmatrix} 1{,}6 & 1{,}2 & 0{,}4 \\ 1{,}2 & 0{,}5 & 1{,}0 \\ 0{,}4 & 1{,}0 & 0{,}8 \end{bmatrix};$$

$$\text{ж)} \begin{bmatrix} 1{,}2 & 0{,}4 & 2{,}5 \\ 0{,}4 & 1{,}4 & 0{,}6 \\ 2{,}5 & 0{,}6 & 0{,}7 \end{bmatrix}; \qquad \text{з)} \begin{bmatrix} 0{,}7 & 0{,}8 & 1{,}3 \\ 0{,}8 & 2{,}6 & 1{,}3 \\ 1{,}3 & 1{,}3 & 0{,}8 \end{bmatrix};$$

$$\text{и)}\quad \begin{bmatrix} 1{,}2 & 2{,}0 & 0{,}5 \\ 2{,}0 & 1{,}4 & 1{,}7 \\ 0{,}5 & 1{,}7 & 0{,}3 \end{bmatrix}.$$

2.4.3. Частичная проблема собственных значений и собственных векторов матрицы. Степенной метод. Рассмотренный метод вращения решает полную проблему собственных значений и собственных векторов матриц (симметрических) в том смысле, что определяются все собственные значения и собственные векторы.

Зачастую не нужно находить все собственные значения (спектр) и все собственные векторы, а необходимо найти максимальное и минимальное из них. Существует степенной метод определения спектрального радиуса матрицы, т. е. максимального собственного значения матрицы, и соответствующего ему собственного вектора.

Пусть дана матрица A, и пусть ее собственные значения упорядочены по абсолютным величинам:

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|. \tag{2.86}$$

Тогда, выбрав некоторый вектор $y^{(0)}$, например вектор, компоненты которого равны единице $y^{(0)} = (1\ 1\ \ldots\ 1)^T$, можно построить следующую итерационную последовательность:

$$\begin{cases} y^{(1)} = Ay^{(0)}, \\ y^{(2)} = Ay^{(1)} = A^2 y^{(0)}, \\ y^{(3)} = Ay^{(2)} = A^3 y^{(0)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y^{(k)} = Ay^{(k-1)} = A^k y^{(0)}, \\ y^{(k+1)} = A^{k+1} y^{(0)}. \end{cases} \tag{2.87}$$

Здесь в круглых скобках индексов указаны номера итераций, а без скобок — показатели степеней.

Покажем, что алгоритм (2.87), называемый *степенным методом*, можно использовать для получения максимального собственного значения и соответствующего ему собственного вектора матрицы A.

Разложим $y^{(0)}$ по собственным векторам матрицы A:

$$y^{(0)} = \alpha_1 x^1 + \alpha_2 x^1 + \ldots + \alpha_n x^n, \tag{2.88}$$

и умножим слева левую и правую части выражения (2.88) на матрицу A^k, получим

$$y^{(k)} = A^k y^{(0)} = \alpha_1 A^k x^1 + \alpha_2 A^k x^2 + \ldots + \alpha_n A^k x^n. \tag{2.89}$$

Рассмотрим в (2.89) слагаемое с сомножителем $A^k x^1$:

$$Ax^1 = \lambda_1 x^1; \quad A^2 x^1 = A(Ax^1) = \lambda_1 A x^1 = \lambda_1^2 x^1;$$

$$A^3 x^1 = \lambda_1^3 x^1; \quad \ldots; \quad A^k x^1 = \lambda_1^k x^1. \tag{2.90}$$

В соответствии с (2.90) выражение (2.89) можно записать следующим образом:

$$y^{(k)} = A^k y^{(0)} = \alpha_1 \lambda_1^k x^1 + \alpha_2 \lambda_2^k x^2 + \ldots + \alpha_n \lambda_n^k x^n. \tag{2.91}$$

Записав выражение (2.91) для j-го компонента вектора $y^{(k)}$ на k-й и $(k+1)$-й итерациях:

$$y_j^{(k)} = \alpha_1 x_j^1 \lambda_1^k + \alpha_2 x_j^2 \lambda_2^k + \ldots + \alpha_n x_j^n \lambda_n^k,$$

$$y_j^{(k+1)} = \alpha_1 x_j^1 \lambda_1^{k+1} + \alpha_2 x_j^2 \lambda_2^{k+1} + \ldots + \alpha_n x_j^n \lambda_n^{k+1},$$

и разделив y_j^{k+1} на y_j^k, получим

$$\frac{y_j^{(k+1)}}{y_j^{(k)}} = \frac{\beta_{1j}\lambda_1^{k+1} + \beta_{2j}\lambda_2^{k+1} + \ldots + \beta_{nj}\lambda_n^{k+1}}{\beta_{1j}\lambda_1^{k} + \beta_{2j}\lambda_2^{k} + \ldots + \beta_{nj}\lambda_n^{k}},$$

где $\beta_{ij} = \alpha_i x_j^i$, $i = \overline{1,n}$. Разделим числитель и знаменатель последнего выражения на $\beta_{1j}\lambda_1^{k+1}$:

$$\frac{y_j^{(k+1)}}{y_j^{(k)}} = \lambda_1 \frac{1 + \gamma_{2j}\mu_2^{k+1} + \ldots + \gamma_{nj}\mu_n^{k+1}}{1 + \gamma_{2j}\mu_2^{k} + \ldots + \gamma_{nj}\mu_n^{k}}, \tag{2.92}$$

где $\mu_i = \dfrac{\lambda_i}{\lambda_1} < 1$, $i = \overline{2,n}$; $\gamma_{ij} = \dfrac{\beta_{ij}}{\beta_{1j}}$, $i = \overline{2,n}$.

Перейдя в (2.92) к пределу при $k \to \infty$,

$$\lim_{k\to\infty} \frac{y_j^{(k+1)}}{y_j^{(k)}} = \lambda_1,$$

а затем сняв знак предела, получим для ограниченного k

$$\lambda_1 \approx \frac{y_j^{(k+1)}}{y_j^{(k)}}, \tag{2.93}$$

где j — любое из $j = \overline{1,n}$.

С помощью выражения (2.93) приближенно вычисляется максимальное собственное значение λ_1 матрицы A (ее спектральный радиус).

Точное выражение для λ_1 может быть записано следующим образом:

$$\lambda_1 = \frac{y_j^{(k+1)}}{y_j^{(k)}} + O\left(\mu_2\right)^{k+1},$$

где $\mu_2 = \dfrac{\lambda_2}{\lambda_1}$, j— любое из $j = \overline{1,n}$.

Для приближенного вычисления собственного вектора, соответствующего максимальному собственному значению λ_1, положим в выражении (2.91) слагаемые, начиная со второго, равными нулю при большом числе итераций, получим

$$y^{(k)} = A^k y^{(0)} \approx \alpha_1 \lambda_1^k x^1,$$

откуда $x^1 \approx \dfrac{1}{\alpha_1\lambda_1^k} y^{(k)}$.

Поскольку собственные векторы определяются с точностью до постоянной, то число α_1 можно положить равным 1, тогда

$$x^1 \approx \frac{1}{\lambda_1^k} y^{(k)}. \tag{2.94}$$

Таким образом, определяющими выражениями в степенном методе являются алгоритм (2.87) и выражения (2.93) — для собственного значения и (2.94) — для собственного вектора. При выполнении условий (2.86) итерационный процесс сходится к искомому собственному значению λ_1 и соответствующему собственному вектору, причем скорость сходимости определяется отношением $|\lambda_2|/|\lambda_1|$ (чем оно меньше, тем выше скорость сходимости).

В качестве критерия завершения вычислений используется следующее условие:

$$\varepsilon^{(k)} = \left|\lambda_1^{(k)} - \lambda_1^{(k-1)}\right| \leqslant \varepsilon,$$

где ε — задаваемая вычислителем точность расчета.

С помощью степенного метода можно вычислить и минимальное собственное значение матрицы A (внутреннего радиуса спектрального кольца). Для этого достаточно степенной метод применить к обратной матрице A^{-1}, т.е. получить $\left|\lambda_{\max}(A^{-1})\right|$ и взять обратную величину:

$$\lambda_{\min}(A) = \frac{1}{\lambda_{\max}(A^{-1})}.$$

Соответствующий собственный вектор x^n будет

$$x^n \approx (A^{-1})^k y^{(0)} \left[\lambda_{\min}(A)\right]^k, \quad \left(y^{(0)}\right)^T = (11\ \ldots\ 1),$$

или, с точностью до константы,

$$x^n = (A^{-1})^k \cdot y^{(0)}.$$

Конечно, вычисление минимального собственного значения матрицы A требует ее обращения, поэтому этот способ вычисления $\lambda_{\min}(A)$ не распространен широко.

Пример 2.20. Вычислить спектральный радиус матрицы

$$A = \begin{pmatrix} 5 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 3 \end{pmatrix}$$

с точностью $\varepsilon = 0{,}1$.

Решение. В качестве начального приближения собственного вектора возьмем $y^{(0)} = (1\ 1\ 1)^T$. Реализуем итерационный процесс (2.87), (2.93), полагая $j = 1$:

$$y^{(1)} = Ay^{(0)} = (8\ \ 6\ \ 6)^T, \quad \lambda_1^{(1)} = \frac{y_1^{(1)}}{y_1^{(0)}} = \frac{8}{1} = 8;$$

$$y^{(2)} = Ay^{(1)} = (58\ \ 38\ \ 40)^T, \quad \lambda_1^{(2)} = \frac{y_1^{(2)}}{y_1^{(1)}} = \frac{58}{8} = 7{,}25;$$

$$\varepsilon^{(2)} = \left|\lambda_1^{(2)} - \lambda_1^{(1)}\right| = 0{,}75 > \varepsilon;$$

$$y^{(3)} = Ay^{(2)} = (480\ \ 250\ \ 274)^T, \quad \lambda_1^{(3)} = \frac{y_1^{(3)}}{y_1^{(2)}} = \frac{480}{58} = 7{,}034;$$

$$\varepsilon^{(3)} = \left|\lambda_1^{(3)} - \lambda_1^{(2)}\right| = 0{,}216 > \varepsilon;$$

$$y^{(4)} = Ay^{(3)} = (2838\ \ 1682\ \ 1888)^T, \quad \lambda_1^{(4)} = \frac{y_1^{(4)}}{y_1^{(3)}} = \frac{2838}{408} = 6{,}9559;$$

$$\varepsilon^{(4)} = \left|\lambda_1^{(4)} - \lambda_1^{(3)}\right| = 0{,}078 < \varepsilon.$$

Таким образом, полученное на 4-ой итерации значение $\lambda_1^{(4)} = 6{,}9559$ удовлетворяет заданной точности и может быть взято в качестве приближенного значения λ_1. Искомое значение спектрального радиуса $\rho(A) = \max\limits_i |\lambda_i| = |\lambda_1| = 6{,}9559$.

Рассмотренный выше пример наглядно иллюстрирует существенный недостаток алгоритма (2.87), связанный с сильным возрастанием компонентов итерируемого вектора $y^{(k)}$ в ходе итерационного процесса. Видно, что $\left|\dfrac{y_j^{(k)}}{y_j^{(k-1)}}\right| \approx |\lambda_1|$. Во избежание

неограниченного возрастания (при $|\lambda_1| > 1$) или убывания (при $|\lambda_1| < 1$) компонентов $y^{(k)}$ по мере увеличения числа итераций k обычно при проведении компьютерных расчетов применяется степенной метод с нормировкой итерируемого вектора. С этой целью алгоритм (2.87) модифицируется следующим образом:

$$z^{(k)} = Ay^{(k-1)}, \quad \lambda_1^{(k)} = \frac{z_j^{(k)}}{y_j^{(k-1)}}, \quad y^{(k)} = \frac{z^{(k)}}{\|z^{(k)}\|}.$$

При этом в качестве начального приближения $y^{(0)}$ берется вектор с единичной нормой.

Широко распространена также версия степенного метода, использующая скалярные произведения:

$$z^{(k)} = Ay^{(k-1)}, \quad y^{(k)} = \frac{z^{(k)}}{\|z^{(k)}\|}, \quad \lambda_1^{(k)} = \left(y^{(k)},\ Ay^{(k)}\right).$$

ГЛАВА III

ТЕОРИЯ ПРИБЛИЖЕНИЙ

Программа

Исчисление конечных и разделенных разностей. Задача интерполяции, единственность многочленной интерполяции. Интерполяционные многочлены Лагранжа и Ньютона. Погрешность многочленной интерполяции. Сплайн-интерполяция. Вывод кубического сплайна дефекта один. Метод наименьших квадратов: точечный и интегральный. Численное дифференцирование с помощью сглаживающих функций и с помощью отношения конечных разностей. Порядок и уточнение формул численного дифференцирования. Численное интегрирование. Методы прямоугольников, трапеций, Симпсона; их геометрическая интерпретация, погрешность, порядок. Повышение порядка методов численного интегрирования.

В теории приближений изучаются методы приближения функций более простыми, хорошо изученными функциями, методы численного дифференцирования и численного интегрирования. При этом исследуемая приближаемая функция может быть задана как в аналитическом, так и дискретном виде (в виде экспериментальной таблицы).

Пусть дана некоторая функция $f(x)$ на отрезке $x \in [a, b]$, которая является довольно сложной для исследования. Требуется заменить эту функцию некоторой простой, но хорошо исследуемой функцией (например, многочленом). Для этого с помощью $f(x)$ строят таблицу (ее называют *сеточной* функцией),

x_i	x_0	x_1	...	x_n
y_i	y_0	y_1	...	y_n

(3.1)

которую можно заменить (сгладить) простой функцией с контролируемой погрешностью.

Рассмотрим два подхода к такой замене.

1. Пусть приближенная функция, являющаяся многочленом n-й степени,

$$\overline{f}\,(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n, \tag{3.2}$$

где $n+1$ — число узлов в таблице (3.1), с неизвестными параметрами a_i, $i = \overline{0,n}$, так приближает сеточную функцию $f\,(x)$, что

$$y_i = \overline{f}(x_i), \quad i = \overline{0,n}. \tag{3.3}$$

В этом случае говорят, что функция $\overline{f}\,(x)$ *интерполирует* сеточную функцию (3.1), а сама задача приближения называется *задачей интерполяции*. Точки x_i, $i = \overline{0,n}$, называют *узлами интерполяции*, а условие (3.3) — *условием интерполяции*. Появляется возможность вычислить значения $\overline{f}\,(x)$ не только в узлах интерполяции, но и между ними в точках $\xi \in (x_{i-1}, x_i)$, $i = \overline{1,n}$, причем $\overline{f}\,(\xi) \approx f\,(\xi)$.

2. При большом количестве точек x_i, $i = \overline{0,n}$, интерполяция требует большой гладкости (по n-й производной), что практически выполнить невозможно. Поэтому сглаживание сеточной функции (3.1) осуществляют путем минимизации некоторого функционала, построенного с помощью (3.1) и многочлена (3.2) степени m, например квадратичного функционала:

$$S\,(a_0, a_1, \ldots, a_m) = \sum_{i=0}^{n} \left[y_i - \overline{f}\,(x_i)\right]^2, \quad m \ll n. \tag{3.4}$$

Процедуру сглаживания в этом случае называют *аппроксимацией* заданной функции функцией (3.2), в частности, аппроксимацию с использованием функционала (3.4) называют аппроксимацией с помощью *точечного метода наименьших квадратов*.

Если коэффициенты сглаживающей функции (3.2) определяются путем минимизации функционала,

$$S\,(a_0, a_1, \ldots, a_m) = \int\limits_a^b \left[f\,(x) - \overline{f}\,(x)\right]^2 dx,$$

сглаживание называют *интегральным методом наименьших квадратов*.

Если в качестве сглаживаемой функции задана экспериментальная таблица (3.1), то в методах сглаживания практически ничего не изменяется. Изменяются методы оценки погрешности сглаживания.

При построении методов сглаживания очень часто используется понятие конечных разностей.

§ 3.1. Исчисление конечных разностей

Пусть дана сеточная функция (3.1) для функции $f(x)$ или экспериментальная таблица (3.1).

В вычислительной математике аналогом понятия дифференциала является понятие конечной разности, которое используется при построении методов теории приближений, в частности при построении интерполяционных многочленов.

Определение 1. Конечной разностью первого порядка вперед (назад) сеточной функции (3.1) в узле x_i называют разность

$$\Delta y_i = y_{i+1} - y_i, \quad i = \overline{0, n-1} \quad \left(\Delta \overline{y}_i = y_i - y_{i-1}, \quad i = \overline{1, n}\right).$$

Определение 2. Конечной разностью второго порядка вперед (назад) сеточной функции (3.1) в узле x_i называют разность первого порядка от разности первого порядка:

$$\Delta^2 y_i = \Delta\left(\Delta y_i\right) = \Delta\left(y_{i+1} - y_i\right) = \Delta y_{i+1} - \Delta y_i = \\ = \left(y_{i+2} - y_{i+1}\right) - \left(y_{i+1} - y_i\right) = y_{i+2} - 2y_{i+1} + y_i,$$

$$i = \overline{0, n-2} \quad \left(\Delta^2 \overline{y}_i = y_{i+1} - 2y_i + y_{i-1}, \quad i = \overline{1, n-1}\right).$$

Определение 3. Конечной разностью 3-го порядка вперед (назад) называется разность первого порядка от разности второго порядка:

$$\Delta^3 y_i = y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i, \quad i = \overline{0, n-3}$$

$$\left(\Delta^3 \overline{y}_i = y_{i+2} - 3y_{i+1} + 3y_i - y_{i-1}, \quad i = \overline{1, n-2}\right)$$

и т. д.

По аналогии запишем конечную разность k-го порядка вперед (назад) в узле x_i:

$$\Delta^k y_i = y_{i+k} - C_k^1 y_{i+k-1} + C_k^2 y_{i+k-2} - \ldots \\ \ldots + (-1)^k y_i, \quad i = \overline{0, n-k}$$

$$\left(\Delta^k \overline{y}_i = y_{i+k-1} - C_k^1 y_{i+k-2} + C_k^2 y_{i+k-3} - \ldots \right. \\ \left. \ldots + (-1)^k y_{i-1}, \quad i = \overline{1, n-k+1}\right). \tag{3.5}$$

Здесь $C_k^m = \dfrac{k\,(k-1)\,(k-2)\ldots(k-m+1)}{m!}$.

Ясно, что конечной разностью сеточной функции (3.1) *нулевого порядка* в узлах x_i, $i=\overline{0,n}$, являются значения y_i этой функции в этих узлах.

§ 3.2. Задача интерполяции

Пусть на отрезке $x \in [a,b]$ задана функция $f(x)$, с помощью которой построена сеточная функция (3.1) или задана экспериментальная таблица (3.1).

При сглаживании функции (или экспериментальной таблицы) с помощью интерполяции в соответствии с условием интерполяции (3.3) значение интерполирующей функции и значение заданной функции в узлах сетки должны быть одинаковыми, следовательно, погрешность интерполяции в узлах x_i, $i=\overline{0,n}$, равна нулю (рис. 3.1).

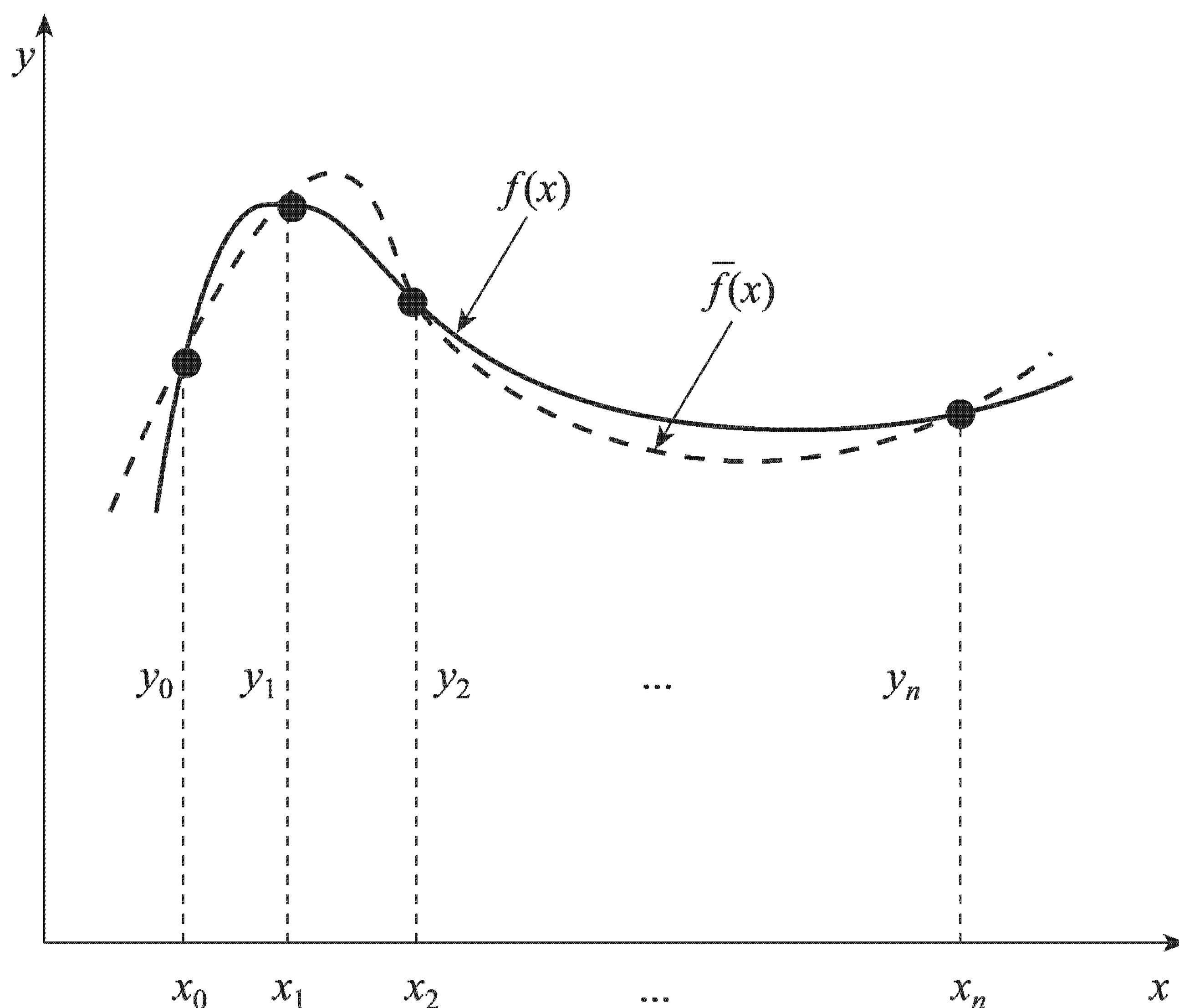


Рис. 3.1. К задаче интерполяции

Задача интерполяции имеет не единственное решение, но в одном случае, когда интерполирующей функцией является многочлен n-й степени ($n+1$ — число узлов интерполяции)

вида (3.2), интерполяция имеет единственное решение, т. е. коэффициенты $a_0, \ldots, a_n$ определяются единственным образом.

Действительно, используя таблицу (3.1) и многочлен (3.2), составим СЛАУ относительно неизвестных коэффициентов $a_0, \ldots, a_n$:

$$\begin{array}{c} i=0 \\ i=1 \\ \ldots \\ i=n \end{array} \left\{ \begin{array}{l} a_0 x_0^n + a_1 x_0^{n-1} + \ldots + a_n = y_0, \\ a_0 x_1^n + a_1 x_1^{n-1} + \ldots + a_n = y_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 x_n^n + a_1 x_n^{n-1} + \ldots + a_n = y_n. \end{array} \right. \tag{3.6}$$

Неоднородная СЛАУ (3.6) имеет единственное решение для коэффициентов $a_0, \ldots, a_n$, так как определитель матрицы этой СЛАУ не равен нулю:

$$\det \begin{pmatrix} x_0^n & x_0^{n-1} & \ldots & 1 \\ x_1^n & x_1^{n-1} & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ x_n^n & x_n^{n-1} & \ldots & 1 \end{pmatrix} \neq 0,$$

поскольку все значения узлов интерполяции различны между собой и ни одна из строк не является линейной комбинацией других. Таким образом, *задача многочленной интерполяции имеет единственное решение*, так как коэффициенты $a_0, \ldots, a_n$ могут быть выбраны единственным образом.

3.2.1. Интерполяционный многочлен Лагранжа. Для многочленной интерполяции можно и не решать СЛАУ (3.6), а многочлен (3.2) можно составить следующим образом.

Запишем систему многочленов n-й степени:

$$l_0 = \frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)} = \begin{cases} 1, & x = x_0, \\ 0, & x = x_i,\ i = \overline{1, n}; \end{cases}$$

$$l_1 = \frac{(x-x_0)(x-x_2)\ldots(x-x_n)}{(x_1-x_0)(x_1-x_2)\ldots(x_1-x_n)} = \begin{cases} 1, & x = x_1, \\ 0, & x = x_i,\ i = 0, \overline{2,\ n}; \end{cases}$$

....

$$l_n = \frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})} = \\ = \begin{cases} 1, & x = x_n, \\ 0, & x = x_i, \ i = \overline{0, n-1}. \end{cases}$$

Составим линейную комбинацию этих многочленов (их количество равно $n+1$) с коэффициентами линейной комбинации, равными значениям y_i сеточной функции (3.1), получим многочлен n-й степени

$$L_n(x) = \\ = \sum_{i=0}^{n} y_i \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}. \tag{3.7}$$

Многочлен (3.7) называют *интерполяционным многочленом Лагранжа* n-й степени, так как он, во-первых, удовлетворяет условию интерполяции

$$L_n(x_i) = y_i, \quad i = \overline{0, n},$$

и, во-вторых, имеет n-ю степень.

Интерполяционный многочлен Лагранжа обладает тем недостатком, что в случае, когда добавляются новые узлы интерполяции в таблице (3.1), все слагаемые в (3.7) необходимо пересчитывать. Но, с другой стороны, он обладает тем достоинством, что интервалы между узлами могут быть неравномерными: $x_{i+1} - x_i = h_i \neq \text{const}$, $i = \overline{0, n-1}$.

Выпишем наиболее употребляемые многочлены $L_1(x)$ и $L_2(x)$.

1) Для таблицы с двумя узлами интерполяции x_i, x_{i+1}:

x_i	x_{i+1}
y_i	y_{i+1}

$$L_1(x) = y_i \frac{x - x_{i+1}}{x_i - x_{i+1}} + y_{i+1} \frac{x - x_i}{x_{i+1} - x_i};$$

2) Для таблицы с тремя узлами интерполяции x_{i-1}, x_i, x_{i+1}:

x_{i-1}	x_i	x_{i+1}
y_{i-1}	y_i	y_{i+1}

$$L_2(x) = y_{i-1}\frac{(x-x_i)(x-x_{i+1})}{(x_{i-1}-x_i)(x_{i-1}-x_{i+1})} + \\ + y_i\frac{(x-x_{i-1})(x-x_{i+1})}{(x_i-x_{i-1})(x_i-x_{i+1})} + y_{i+1}\frac{(x-x_{i-1})(x-x_i)}{(x_{i+1}-x_{i-1})(x_{i+1}-x_i)}.$$

3.2.2. Интерполяционный многочлен Ньютона. Пусть сеточная функция (3.1) имеет равномерный шаг между узлами интерполяции: $x_{i+1} - x_i = h_i = h = \text{const}$, $i = \overline{0, n-1}$.

Будем строить интерполяционный многочлен следующим образом:

$$N_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \ldots \\ \ldots + a_n(x-x_0)\ldots(x-x_{n-1}). \quad (3.8)$$

Коэффициенты $a_0, \ldots, a_n$ будем определять из условия интерполяции $N_n(x_i) = y_i$, $i = \overline{0, n}$, с использованием таблицы (3.1):

$$x = x_0 : N_n(x_0) = y_0 = a_0 \Rightarrow a_0 = y_0; \quad (3.9)$$

$$x = x_1 : N_n(x_1) = a_0 + a_1(x_1 - x_0) = \\ = y_1 \Rightarrow a_1 = \frac{y_1 - y_0}{x_1 - x_0}, \quad a_1 = \frac{\Delta y_0}{h}; \quad (3.10)$$

$$x = x_2 : N_n(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = \\ = y_2 \Rightarrow a_2 = \frac{y_2 - 2y_1 - y_0}{2h^2} = \frac{\Delta^2 y_0}{2!h^2}, \quad a_2 = \frac{\Delta^2 y_0}{2!h^2}. \quad (3.11)$$

И так далее,

$$a_n = \frac{\Delta^n y_0}{n!h^n}. \quad (3.12)$$

Подставляя коэффициенты (3.9)–(3.12) в многочлен (3.8), получим

$$N_n(x) = y_0 + \frac{\Delta y_0}{1!h}(x-x_0) + \frac{\Delta^2 y_0}{2!h^2}(x-x_0)(x-x_1) + \ldots \\ \ldots + \frac{\Delta^n y_0}{n!h^n}(x-x_0)\ldots(x-x_{n-1}). \quad (3.13)$$

Многочлен (3.13) является интерполяционным многочленом, поскольку является многочленом n-й степени и удовлетворяет условию интерполяции (3.3). Он называется *интерполяционным многочленом Ньютона*. Его достоинство заключается в том, что

он строится проще, чем $L_n(x)$, и при добавлении новых узлов интерполяции в таблицу (3.1) все предыдущие слагаемые не пересчитываются, а добавляются новые. К недостаткам многочлена $N_n(x)$ по сравнению с $L_n(x)$, можно отнести использование постоянного шага между узлами интерполяции (ниже, в п. 3.2.4, на основе понятия разделенных разностей строится интерполяционный многочлен Ньютона с переменным шагом).

В силу единственности многочленной интерполяции ясно, что после раскрытия скобок получим $L_n(x) = N_n(x)$.

3.2.3. Погрешность многочленной интерполяции. Ясно, что в узлах интерполяции погрешность интерполяционного многочлена $L_n(x)$ или $N_n(x)$ равна нулю:

$$\left\{\begin{array}{c} L_n(x_i) - y_i \\ \\ N_n(x_i) - y_i \end{array}\right\} = 0, \quad i = \overline{0, n}.$$

Будем искать погрешность $L_n(x) - f(x)$ в виде разности между значением интерполяционного многочлена $L_n(x)$ и значением функции $f(x)$ в точке x, не совпадающей с узлом интерполяции.

Для нахождения погрешности составим следующую вспомогательную функцию:

$$\varphi(x) = f(x) - L_n(x) - \alpha \cdot \omega(x), \tag{3.14}$$

где α подлежит определению, а $\omega(x)$ — многочлен $(n+1)$-й степени:

$$\omega(x) = (x - x_0)(x - x_1)\ldots(x - x_n). \tag{3.15}$$

Из (3.14) и (3.15) видно, что $[L_n(x)]^{(n+1)} = 0$, $[\omega(x)]^{(n+1)} = (n + 1!)$.

Будем искать α из условия, что $\varphi(x) = 0$ в точке, в которой исследуется погрешность. Обозначим эту точку через $\overline{x}$:

$$\alpha = \frac{f(\overline{x}) - L_n(\overline{x})}{\omega(\overline{x})}. \tag{3.16}$$

Функция $\varphi(x)$ имеет $(n+1)$ корень в узлах интерполяции, так как там $\omega(x) = 0$ и погрешность $f(x) - L_n(x) = 0$. Но если добавим еще точку $\overline{x}$, в которой потребуем $\varphi(\overline{x}) = 0$, то $\varphi(x)$ будет иметь уже $(n+2)$ корня. Итак $\varphi(x)$ имеет $(n+2)$ корня, $\varphi'(x)$ имеет $(n+1)$ корень, $\varphi''(x)$ имеет n корней и т. д.,

$\varphi^{(n+1)}(x)$ имеет 1 корень. Обозначим этот корень через $\xi \in (x_0, x_n)$.

Покажем, что увеличение порядка производной функции $\varphi(x)$ на единицу уменьшает количество корней этой функции на единицу. Действительно, из теоремы Ролля известно, что если функция $\varphi(x)$ на $[x_{i-1}, x_i]$ непрерывна, дифференцируема на интервале (x_{i-1}, x_i) и имеет на концах одинаковые значения, то внутри этого интервала найдется хотя бы одна точка, в которой производная этой функции равна нулю. Функция $\varphi(x)$ на каждом отрезке удовлетворяет теореме Ролля (см. рис. 3.2).

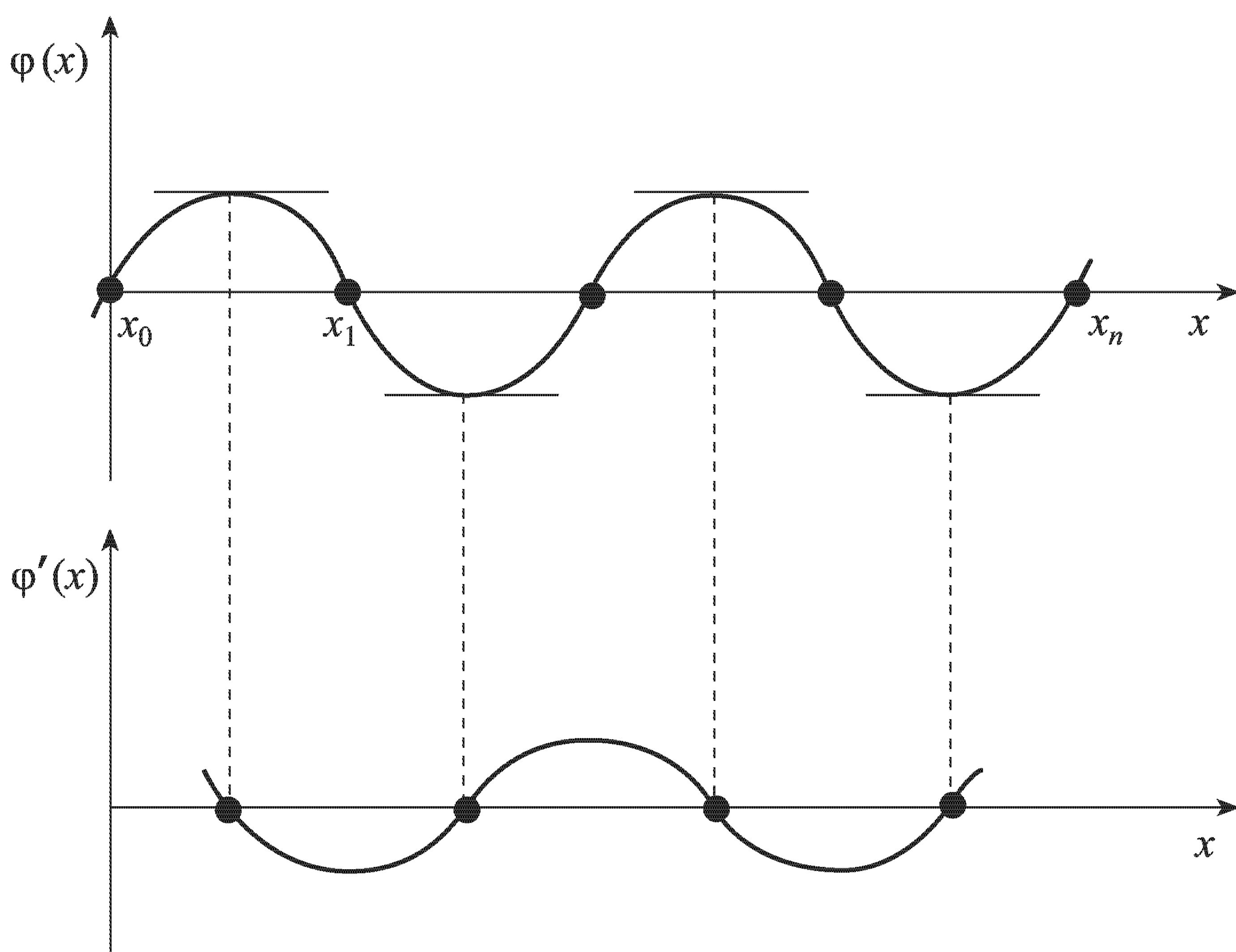


Рис. 3.2. Соотношение количества нулей функций $\varphi(x)$ и $\varphi'(x)$

Вычисляя производную $(n+1)$-го порядка от (3.14) в точке ξ и учитывая, что значение $\varphi^{(n+1)}(x) = 0$, так как $\varphi(x)$ имеет $(n+2)$ корня, получим

$$f^{(n+1)}(\xi) - 0 - \alpha \cdot (n+1)! = 0,$$

откуда

$$\alpha = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{3.17}$$

Подставляя (3.17) в (3.16), находим погрешность многочленной интерполяции в точке $\overline{x} \in (x_{i-1}, x_i)$, $i = \overline{1, n}$:

$$f(\overline{x}) - L_n(\overline{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(\overline{x} - x_0)(\overline{x} - x_1)\ldots(\overline{x} - x_n), \qquad \xi \in (x_0, x_n) \equiv (a, b); \tag{3.18}$$

$\overline{x}$ — точка, в которой ищется погрешность, не совпадает с узлами интерполяции.

Поскольку точка $\xi \in (a, b)$ неизвестна, то вместо погрешности (3.18) вводится верхняя оценка погрешности в виде

$$|f(\overline{x}) - L_n(\overline{x})| \leqslant \leqslant \frac{\max\limits_{x \in [a,b]} \left| f^{(n+1)}(x) \right|}{(n+1)!} \left| (\overline{x} - x_0)(\overline{x} - x_1)\ldots(\overline{x} - x_n) \right|, \tag{3.19}$$

которая и используется на практике.

В случае, если интерполяционный многочлен строится для экспериментальной таблицы (3.1), функция $f(x)$ отсутствует и погрешностью в форме (3.18) или (3.19) воспользоваться не удается. Тогда, добавляя к таблице (3.1) один узел интерполяции, можно записать верхнюю оценку погрешности (3.19) с помощью отношения конечных разностей $(n+1)$-го порядка

$$|y(\overline{x}) - L_n(\overline{x})| \leqslant \frac{\left|\Delta^{(n+1)}(y_0)\right|}{h^{n+1}(n+1)!} \left| (\overline{x} - x_0)(\overline{x} - x_1)\ldots(\overline{x} - x_n) \right|,$$

т. е. с помощью добавочного $(n+1)$-го слагаемого в интерполяционном многочлене Ньютона.

3.2.4. Интерполяционный многочлен Ньютона, построенный с помощью разделенных разностей. В вычислительной математике аналогом понятия производной является понятие *разделенной разности*.

Разделенной разностью сеточной функции (3.1) *нулевого порядка* в узлах x_i, $i = \overline{0, n}$, называются значения этой функции в этих узлах $f(x_i) = y_i$, $i = \overline{0, n}$.

Определение 1. Разделенной разностью функции (3.1) *первого порядка* в узлах x_i, $i = \overline{0, n-1}$, называют отношение

$$f(x_i, x_{i+1}) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \quad i = \overline{0, n-1}.$$

Определение 2. Разделенной разностью функции (3.1) *второго порядка* в узлах x_i, $i = \overline{0, n-2}$, называют отношение

$$f(x_i, x_{i+1}, x_{i+2}) = \frac{f(x_{i+1}, x_{i+2}) - f(x_i, x_{i+1})}{x_{i+2} - x_i} =$$

$$= \frac{\dfrac{f(x_{i+2}) - f(x_{i+1})}{x_{i+2} - x_{i+1}} - \dfrac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}}{x_{i+2} - x_i} = \frac{\dfrac{y_{i+2} - y_{i+1}}{x_{i+2} - x_{i+1}} - \dfrac{y_{i+1} - y_i}{x_{i+1} - x_i}}{x_{i+2} - x_i},$$

$i = \overline{0, n-2}$. И так далее.

Определение 3. Разделенной разностью функции (3.1) *n-го порядка* в узле x_0 называют отношение

$$f(x_0, x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n) - f(x_0, \ldots, x_{n-1})}{x_n - x_0}.$$

Если шаг сетки в таблице (3.1) постоянный, т. е. $h_i = x_i - x_{i-1} = \text{const}$, $i = \overline{1, n}$, то разделенные разности совпадают с отношением конечных разностей. Например, разделенные разности 2-го и 3-го порядков в узле x при $x_i - x_{i-1} = \text{const}$ имеют вид

$$f(x_0, x_1, x_2)\big|_{h=\text{const}} = \frac{y_2 - 2y_1 + y_0}{2!h^2} = \frac{1}{2!} \cdot \frac{\Delta^2 y_0}{h^2},$$

$$f(x_0, x_1, x_2, x_3)\big|_{h=\text{const}} =$$

$$= \frac{f(x_1, x_2, x_3) - f(x_0, x_1, x_2)}{x_3 - x_0} = \frac{\dfrac{\Delta^2 y_1}{2!h^2} - \dfrac{\Delta^2 y_0}{2!h^2}}{3h} = \frac{\Delta^3 y_0}{3!h^3}.$$

Таким образом, если в интерполяционном многочлене Ньютона коэффициенты при многочленах в слагаемых заменить разделенными разностями, т. е. многочлен Ньютона записать в форме

$$N_n(x) =$$
$$= f(x_0) + f(x_0, x_1)(x - x_0) + f(x_0, x_1, x_2)(x - x_0)(x - x_1) + \ldots$$
$$\ldots + f(x_0, x_1, \ldots, x_n)(x - x_0)(x - x_1)\ldots(x - x_{n-1}), \quad (3.20)$$

то при $h = \text{const}$ эта форма превращается в обычную выведенную ранее форму (3.13):

$$N_n(x)\big|_{h=\text{const}} = y_0 + \frac{\Delta y_0}{1!h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) + \ldots$$
$$\ldots + \frac{\Delta^n y_0}{n!h^n}(x - x_0)(x - x_1)\ldots(x - x_{n-1}). \quad (3.21)$$

Итак, интерполяционный многочлен Ньютона (3.20) используется в случае задания таблицы (3.1) с неравномерным шагом, а (3.21) — для таблицы (3.1) с постоянным шагом.

3.2.5. Сплайн-интерполяция. Интерполяция, использующая сразу все n узлов таблицы (3.1), называется *глобальной интерполяцией.*

Начиная с $n \geqslant 7$ глобальная многочленная интерполяция становится неустойчивой в том смысле, что погрешности возрастают, так как записанный интерполяционный многочлен требует гладкости по производным 7-го и высших порядков (рис. 3.3).

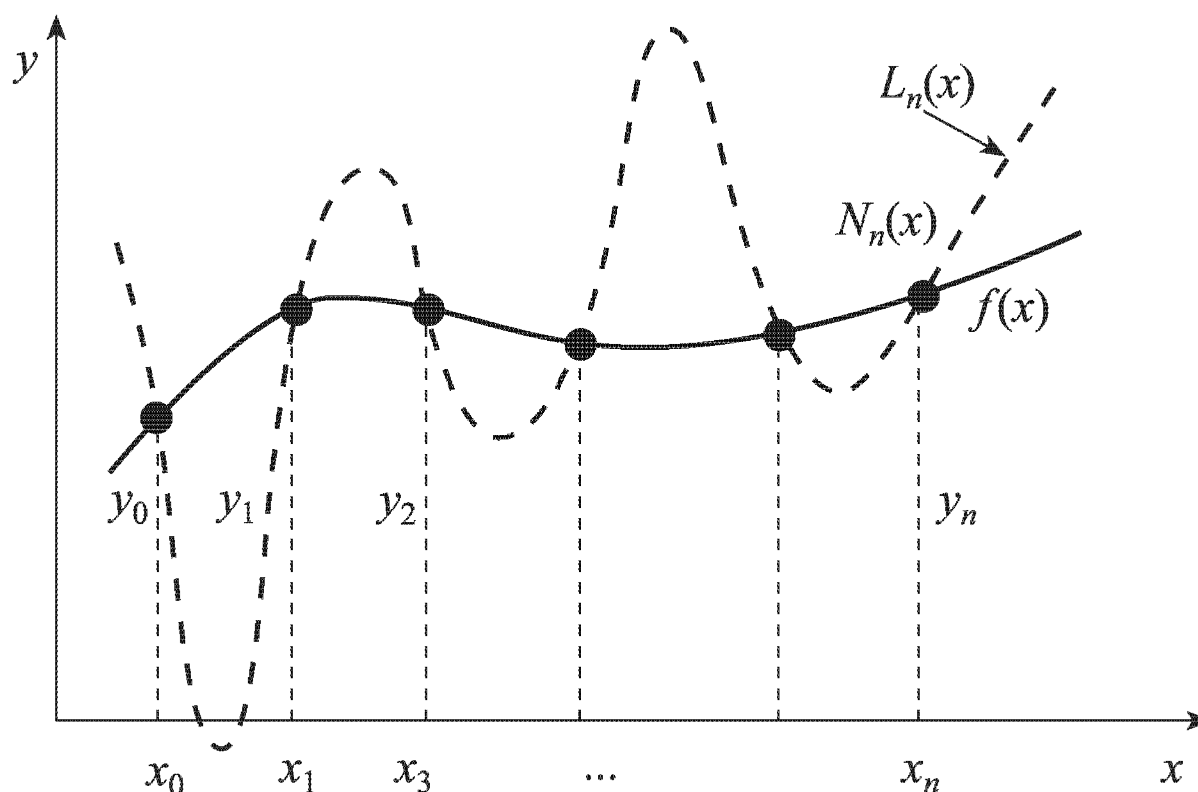


Рис. 3.3. Неустойчивость глобальной многочленной интерполяции при $n \geqslant 7$

Поэтому обычную многочленную интерполяцию осуществляют максимум по 3–4 узлам (для 3-х узлов 2-ой степени, 4-х узлов 3-й степени). Интерполяцию по нескольким узлам таблицы (3.1) называют *локальной*: линейной по каждым двум узлам с помощью интерполяционных многочленов $L_1(x)$, квадратичной по каждым трем узлам с помощью интерполяционных многочленов $L_2(x)$, и так далее.

Однако такая локальная интерполяция с помощью L_n или N_n страдает тем недостатком, что интерполирующая функция в узлах стыковки многочлена имеет непрерывность только нулевого порядка, т. е. локальные интерполяционные многочлены принадлежат классу функций C^0 (см. рис. 3.4 для L_2, N_2 в узле x^*).

От этих недостатков свободна сплайн-интерполяция, которая требует непрерывности в узлах стыковки локальных многочленов по производным соответственно порядка один, два и т. д.

Определение. Сплайном степени m дефекта r называется $(m - r)$ раз непрерывно дифференцируемая функция, которая

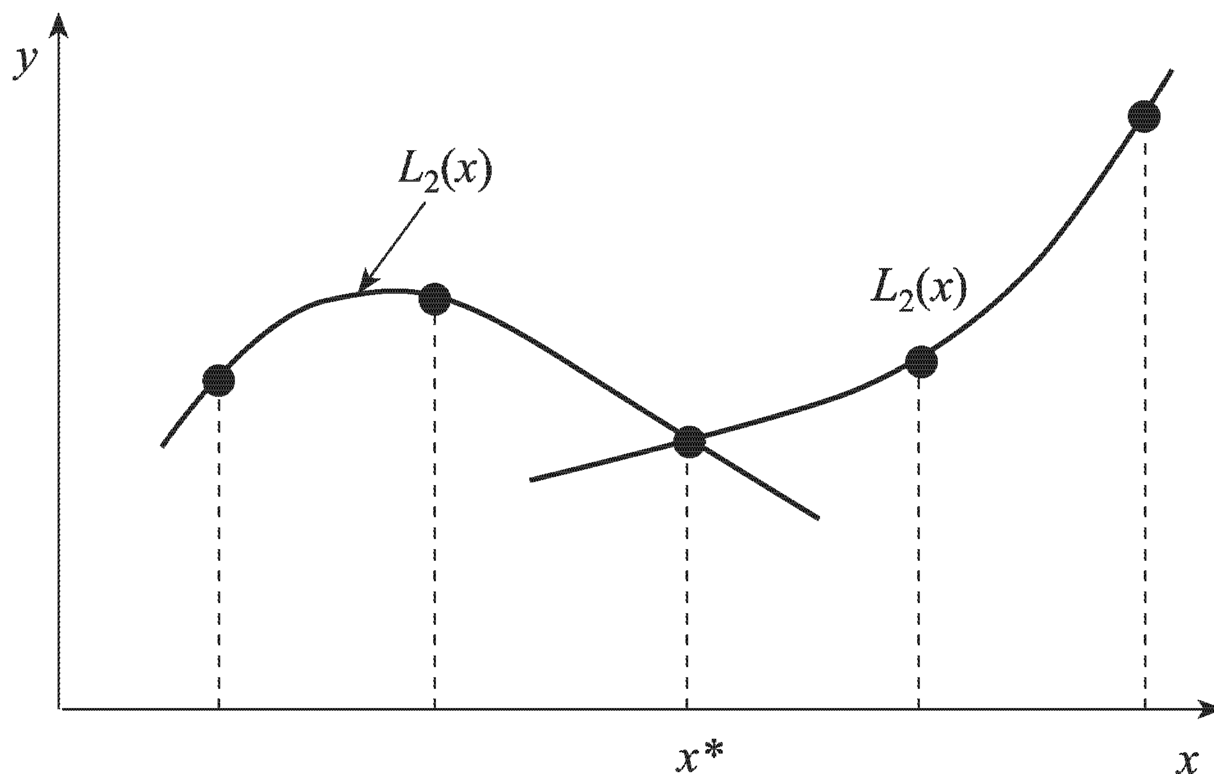


Рис. 3.4. Локальная интерполяция по каждым трем узлам

на каждом отрезке $[x_{i-1}, x_i]$, $i = \overline{1, n}$, представляет собой многочлен степени m.

Наиболее распространенными в науке и технике являются сплайны 3-й степени дефекта один, т. е.

$$\left.\begin{aligned} m &= 3, \\ r &= 1 \end{aligned}\right\} \Rightarrow m - r = 3 - 1 = 2,$$

т. е. дважды непрерывно дифференцируемый многочлен 3-й степени на каждом отрезке $[x_{i-1}, x_i]$, $i = \overline{1, n}$. Сплайны, удовлетворяющие условию интерполяции, называются *интерполяционными*.

Основным достоинством интерполяционного кубического сплайна $S(x)$ дефекта один является следующее: этот сплайн обладает минимумом интегральной кривизны на всем заданном отрезке $[a, b]$ по сравнению с другими интерполяционными функциями $\overline{f}(x)$, т. е.

$$\int\limits_a^b \left[S''(x)\right]^2 dx \leqslant \int\limits_a^b \left[\overline{f}''(x)\right]^2 dx.$$

Геометрически это означает, что если тяжелую упругую нить

повесить на ряд гвоздей, то она примет форму кубического сплайна дефекта 1, приведенную на рис. 3.5.

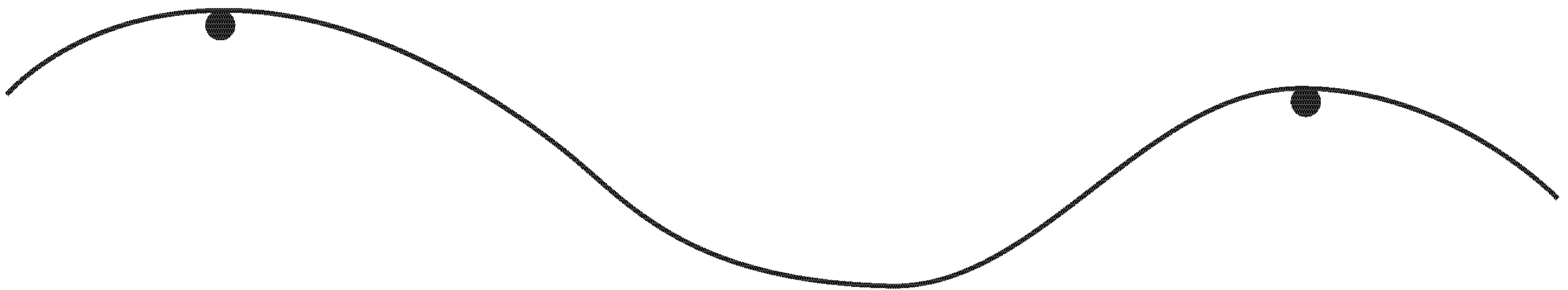


Рис. 3.5. Тяжелая упругая нить, геометрически представляющая собой кубические сплайны дефекта один

Выведем интерполяционные кубические сплайны $S_k(x)$, $k = \overline{1, n}$, дефекта один в соответствии с таблицей (3.1). Кубический сплайн $S(x)$ на отрезке $x \in [x_{i-1}, x_i]$ имеет четыре неизвестных коэффициента. Количество отрезков $[x_{i-1}, x_i]$ в соответствии с таблицей (3.1) равно n. Для определения $4 \times n$ коэффициентов имеются следующие условия в узлах интерполяции:

— условие интерполяции $S(x_i) = y_i$, $i = \overline{0, n}$;

— непрерывность сплайнов $S(x_i - 0) = S(x_i + 0)$, $i = \overline{1, n-1}$;

— непрерывность производных 1-го порядка $S'(x_i - 0) = S'(x_i + 0)$, $i = \overline{1, n-1}$;

— непрерывность производных 2-го порядка $S''(x_i - 0) = S''(x_i + 0)$, $i = \overline{1, n-1}$.

Таким образом, всего имеется $(n+1) + 3(n-1) = 4n - 2$ условий. В качестве двух недостающих условий задают значения производных 1-го или 2-го порядка в узлах x_0 и x_n. Для вывода используем значения $S''(x_0) = S''(x_n) = 0$. В этом случае сплайн называется *естественным*.

Пусть $S''(x) = q(x)$. На отрезке $[x_{i-1}, x_i]$ рассмотрим поведение функции $q(x)$ (см. рис. 3.6).

Поскольку сплайн является многочленом 3-й степени, то на каждом отрезке $[x_{i-1}, x_i]$ 2-я производная будет линейна. Найдем ее с помощью интерполяционного многочлена Лагранжа 1-й степени $L_1(x)$:

$$q(x) = q_{i-1}\frac{x - x_i}{x_{i-1} - x_i} + q_i\frac{x - x_{i-1}}{x_i - x_{i-1}}. \tag{3.22}$$

Выражение (3.22) уже удовлетворяет условиям непрерывности производных 2-го порядка. Действительно, подставим в (3.22) $x = x_i - 0$, получим $q(x_i - 0) = q_i$. Затем, выписывая выражение (3.22) для отрезка $[x_i, x_{i+1}]$:

$$q(x) = q_i \frac{x_{i+1} - x}{h_{i+1}} + q_{i+1} \frac{x - x_i}{h_{i+1}}, \quad x \in [x_i, x_{i+1}], \quad i = \overline{1, n-1},$$

и подставляя в него $x_i + 0$ вместо x, получим $q(x_i + 0) = q_i$, что и требовалось показать.

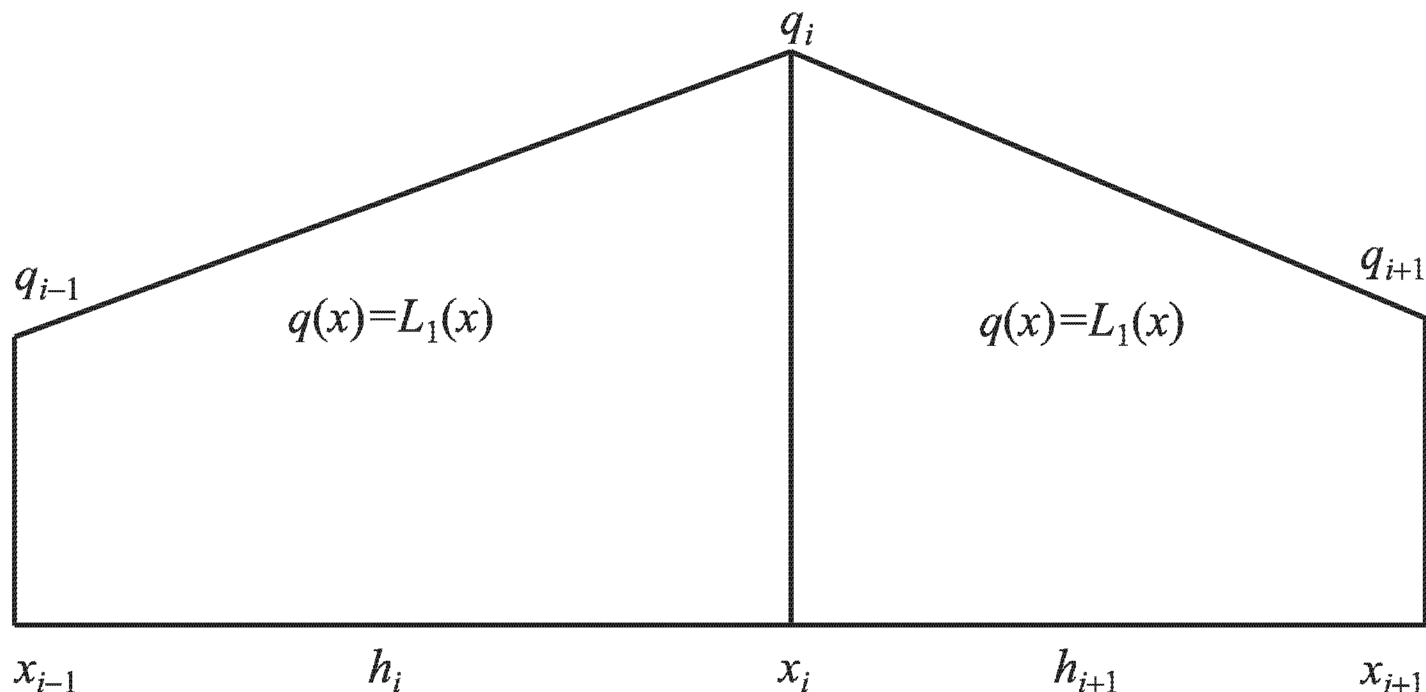


Рис. 3.6. Поведение функций $S''(x)$ на элементарных отрезках

Для нахождения сплайна проинтегрируем дважды выражение (3.22), получим

$$S(x) = q_{i-1} \frac{(x_i - x)^3}{6h_i} + q_i \frac{(x - x_{i-1})^3}{6h_i} + C_1 x + C_2, \tag{3.23}$$

где C_1 и C_2 найдем из удовлетворения значений сплайна (3.23) в узлах x_{i-1}, x_i, условиям интерполяции

$$\begin{cases} S(x_{i-1}) = \\ = y_{i-1} = q_{i-1} \dfrac{(x_i - x_{i-1})^3}{6h_i} + q_i \dfrac{(x_{i-1} - x_{i-1})^3}{6h_i} + C_1 x_{i-1} + C_2, \\ S(x_i) = y_i = q_{i-1} \dfrac{(x_i - x_i)^3}{6h_i} + q_i \dfrac{(x_i - x_{i-1})^3}{6h_i} + C_1 x_i + C_2. \end{cases}$$

Решая эту СЛАУ относительно C_1, C_2 и подставляя их в (3.23), найдем следующее выражение для сплайна степени 3 дефекта 1:

$$S(x) = q_{i-1} \frac{(x_i - x)^3}{6h_i} + q_i \frac{(x - x_{i-1})^3}{6h_i} + \left(\frac{y_{i-1}}{h_i} - q_{i-1} \frac{h_i}{6} \right) \times$$
$$\times (x_i - x) + \left(\frac{y_i}{h_i} - q_i \frac{h_i}{6} \right) (x - x_{i-1}), \quad x \in [x_{i-1}, x_i]. \tag{3.24}$$

В этом сплайне узловые значения для вторых производных q_i пока неизвестны. Будем искать их из условий непрерывности первых производных в узлах x_i.

Для нахождения производной $S'(x_i + 0)$ запишем (3.24) для отрезка $[x_i, x_{i+1}]$

$$S(x) = q_i \frac{(x_{i+1} - x)^3}{6h_{i+1}} + q_{i+1} \frac{(x - x_i)^3}{6h_{i+1}} + \left(\frac{y_i}{h_{i+1}} - q_i \frac{h_{i+1}}{6}\right)(x_{i+1} - - x) + \left(\frac{y_{i+1}}{h_{i+1}} - q_{i+1} \frac{h_{i+1}}{6}\right)(x - x_i), \quad x \in [x_i, x_{i+1}]. \quad (3.25)$$

Вычисляя производные первого порядка от (3.24) и (3.25) и подставляя в них значение $x = x_i$, получим

$$S'(x_i - 0) = q_{i-1} \frac{h_i}{6} + q_i \frac{h_i}{3} + \frac{y_i - y_{i-1}}{h_i},$$

$$S'(x_i + 0) = -q_i \frac{h_{i+1}}{3} - q_{i+1} \frac{h_{i+1}}{6} + \frac{y_{i+1} - y_i}{h_{i+1}}.$$

Приравняем эти выражения в соответствии с условиями непрерывности первых производных в узлах интерполяции x_i, получим

$$q_{i-1} \frac{h_i}{6} + q_i \frac{h_i + h_{i+1}}{3} + q_{i+1} \frac{h_{i+1}}{6} = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}, \quad i = \overline{1, n-1}; \quad (3.26)$$

$$q_0 = q_n = 0. \quad (3.27)$$

Система (3.26) с заданными краевыми условиями (3.27) — СЛАУ относительно $q_i = S''(x_i)$, $i = \overline{1, n-1}$, имеет трехдиагональную матрицу, и, следовательно, ее можно решать методом прогонки. Подставляя найденные q_i, $i = \overline{0, n}$, в (3.24), получим кубические сплайны дефекта один на каждом отрезке $x \in \in [x_{i-1}, x_i]$, $i = \overline{1, n}$.

Таким образом, определяющими выражениями для нахождения кубических сплайнов дефекта один являются выражения (3.24), (3.26), (3.27).

Пример 3.1. Построить интерполяционные многочлены Лагранжа и Ньютона, совпадающие с функцией $f(x) = 3^x$, $x \in \in [-1, 1]$, в точках $x_0 = -1$, $x_1 = 0$, $x_2 = 1$. Вычислить значение

сеточной функции и оценить погрешность многочленной интерполяции в точке $x^* = 0{,}5$.

Решение. Составим сеточную функцию и занесем ее в таблицу. Поскольку $n = 2$, то необходимо построить интерполяционные многочлены $L_2(x)$ и $N_2(x)$:

x_i	$x_0 = -1$	$x_1 = 0$	$x_2 = 1$
y_i	$y_0 = 1/3$	$y_1 = 1$	$y_2 = 3$

$$L_2(x) = y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + \\ + y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{2}{3}x^2 + \frac{4}{3}x + 1.$$

Проверим условия интерполяции $L_2(-1) = 1/3$; $L_2(0) = 1$; $L_2(1) = 3$.

Для определения конечных разностей, входящих в интерполяционный многочлен Ньютона, удобно пользоваться таблицей конечных разностей, в которую конечная разность k-го порядка в узле x_i определяется как $\Delta^{(k)} y_i = \Delta^{(k-1)} y_{i+1} - \Delta^{(k-1)} y_i$, $k = 1, 2, \ldots$; $i = \overline{0, n}$.

x_i	y_i	Δy_i	$\Delta^2 y_i$
-1	$\underline{1/3}$	$1 - 1/3 = \underline{2/3}$	$2 - 2/3 = \underline{4/3}$
0	1	$3 - 1 = 2$	
1	3		

Тогда

$$N_2(x) = y_0 + \frac{\Delta y_0}{1!h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) = \\ = \frac{1}{3} + \frac{2/3}{1 \cdot 1}(x + 1) + \frac{4/3}{2!1^2}(x + 1)(x - 0) = \frac{2}{3}x^2 + \frac{4}{3}x + 1.$$

Здесь использованы конечные разности в узле x_0 (подчеркнуты в таблице). Получен результат, подтверждающий теорему о единственности многочленной интерполяции, поскольку $L_2(x) \equiv N_2(x)$.

Значение сеточной функции в точке $x^* = 0{,}5$ вычислим по интерполяционному многочлену $y\,(0{,}5) \approx L_2\,(0{,}5) = 1{,}8333$.

Верхнюю оценку погрешности интерполяционного многочлена определим в соответствии с выражением (3.19):

$$|f\,(x^*) - L_2\,(x^*)| \leqslant \frac{\max\limits_{x\in[-1,\,1]} |f'''\,(x)|}{3!}\,|(x^* - x_0)\,(x^* - x_1)\,(x^* - x_2)|\,;$$

$$\max_{x\in[-1,\,1]} \left|f'''\,(x)\right| = \max_{x\in[-1,\,1]} \left|3^x \cdot \ln^3 3\right| = 3^1 \cdot \ln^3 3 = 3{,}978;$$

$$\left|3^{0,5} - \left(\frac{2}{3}\cdot 0{,}25 + \frac{4}{3}\cdot 0{,}5 + 1\right)\right| \leqslant$$
$$\leqslant \frac{3{,}978}{6}\,|(0{,}5+1)\,(0{,}5-0)\,(0{,}5-1)| = 0{,}249.$$

Поскольку функция $f(x) = 3^x$ известна, то можно вычислить точное значение абсолютной погрешности в точке $x^* = 0{,}5$:

$$\left|3^{x^*} - L_2\,(x^*)\right| = \left|3^{0,5} - \left(\frac{2}{3}\cdot 0{,}25 + \frac{4}{5}\cdot 0{,}5 + 1\right)\right| = 0{,}1012,$$

т. е. верхняя оценка погрешности примерно в 2,5 раза превышает абсолютную погрешность в точке $x^* = 0{,}5$.

Пример 3.2. Оценить погрешность в точке $x^* = 0{,}5$ многочленной интерполяции из примера 3.1 для случая, когда заданная таблица получена из эксперимента (функция $f\,(x)$ отсутствует).

Решение. В этом случае верхнюю оценку погрешности можно оценить с помощью $(n+1)$-го члена формулы (3.20)или (3.21), дополнив таблицу еще одним узлом интерполяции со значением функции, полученным с помощью, например, экстраполяции.

	x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$
	-1	$1/3$			
	0	1	$2/3$		
	1	3	2	$4/3$	
Доп. узел	2	7	4	2	$2/3$

$$|y\,(x^*) - L_3\,(x^*)| \leqslant \frac{\Delta^3 y_0}{3!h^3}\,|(x^* - x_0)\,(x^* - x_1)\,(x^* - x_2)| =$$

$$= \frac{2/3}{3!\,1^3}\left|(0{,}5+1)\,(0{,}5-0)\,(0{,}5-1)\right| = 0{,}125,$$

т. е. получили погрешность, близкую к истинной абсолютной погрешности, равной 0,1012.

Пример 3.3. Для заданной таблицы

x_i	$x_0=-1$	$x_1=0$	$x_2=2$
y_i	$y_0=1/3$	$y_1=1$	$y_2=9$

составить интерполяционный многочлен Ньютона.

Решение. Таблица задана с неравномерным шагом, поэтому для решения задачи воспользуемся многочленом Ньютона с разделенными разностями (формула (3.20)):

$$N_2(x) = f(x_0) + f(x_0,x_1)(x-x_0) + \\ + f(x_0,x_1,x_2)(x-x_0)(x-x_1),$$

где $f(x_0)=y_0=1/3$;

$$f(x_0,x_1) = \frac{y_1-y_0}{x_1-x_0} = \frac{1-1/3}{0+1} = \frac{2}{3};$$

$$f(x_0,x_1,x_2) = \frac{\dfrac{y_2-y_1}{x_2-x_1}-\dfrac{y_1-y_0}{x_1-x_0}}{x_2-x_0} = \frac{\dfrac{9-1}{2-0}-\dfrac{1-1/3}{0+1}}{2+1} = \frac{10}{9}.$$

Таким образом, $N_2(x) = \dfrac{10}{9}x^2 + \dfrac{16}{9}x + 1.$

Условия интерполяции соблюдены: $N_2(-1)=1/3$; $N_2(0)=1$; $N_2(2)=9$.

Пример 3.4. Для заданной таблицы с $h = x_i - x_{i-1} = 1 = \mathrm{const}$ выписать интерполяционные кубические сплайны дефекта один на каждом отрезке $x\in[x_{i-1},x_i]$, $i=\overline{1,\,4}$. Проверить непрерывность сплайнов и их производных до второго порядка включительно в узле $x^*=2$.

i	0	1	2	3	4
x_i	$x_0=1$	$x_1=2$	$x_2=3$	$x_3=4$	$x_4=5$
y_i	$y_0=1$	$y_1=3$	$y_2=6$	$y_3=9$	$y_4=21$
q_i	0	18/7	−30/7	102/7	0

Решение. Под заданной таблицей сформируем дополнительную строку для вторых производных сплайнов $S''(x_i) \equiv \equiv q_i$, которая заполняется по мере их вычисления (сразу можно вписать в нее $q_0 = q_4 = 0$).

Для узлов $x_1 = 2$; $x_2 = 3$; $x_3 = 4$ с учетом $q_0 = q_4 = 0$ составляется СЛАУ (3.26) относительно неизвестных q_1, q_2, q_3:

$$\begin{aligned} i=1: \\ \\ i=2: \\ \\ i=3: \end{aligned} \left\{ \begin{aligned} &\frac{2}{3}q_1 + \frac{1}{6}q_2 = \frac{y_2 - y_1}{1} - \frac{y_1 - y_0}{1} = 1, \\ &\frac{1}{6}q_1 + \frac{2}{3}q_2 + \frac{1}{6}q_3 = \frac{y_3 - y_2}{1} - \frac{y_2 - y_1}{1} = 0, \\ &\frac{1}{6}q_2 + \frac{2}{3}q_3 = \frac{y_4 - y_3}{1} - \frac{y_3 - y_2}{1} = 9. \end{aligned} \right.$$

Вычисляются прогоночные коэффициенты по формулам

$$A_i = \frac{-c_i}{b_i + a_i A_{i-1}}; \quad B_i = \frac{d_i - a_i B_{i-1}}{b_i + a_i A_{i-1}}, \quad i = 1, 2, 3;$$

$$a_1 = c_3 = 0, \qquad A_1 = -\frac{1}{4}; \;\; B_1 = \frac{3}{2}; \quad A_2 = -\frac{4}{15}; \;\; B_2 = -\frac{2}{5};$$

$$A_3 = 0; \;\; B_3 = \frac{102}{7}$$

и значения

$$q_i = A_i q_{i+1} + B_i, \;\; i = 3, 2, 1: \;\; q_3 = A_3 q_4 + B_3 = B_3 = 102/7;$$

$$q_2 = A_2 q_3 + B_2 = -\frac{4}{15} \cdot \frac{102}{7} - \frac{2}{5} = -\frac{30}{7};$$

$$q_1 = A_1 q_2 + B_1 = -\frac{1}{4} \cdot \left(-\frac{30}{7}\right) + \frac{3}{2} = \frac{18}{7}.$$

Заносим эти значения в дополнительную строку таблицы и для каждого из четырех интервалов выписываем сплайны (3.24):

$$\begin{aligned} i = 1: \; S_I(x) &= q_0 \frac{(x_1 - x)^3}{6 \cdot 1} + q_1 \frac{(x - x_0)^3}{6 \cdot 1} + \\ &+ \left(\frac{y_0}{1} - q_0 \cdot \frac{1}{6}\right)(x_1 - x) + \left(\frac{y_1}{1} - q_1 \cdot \frac{1}{6}\right)(x - x_0) = \\ &= \frac{18}{42}(x-1)^3 + (2 - x) + \frac{108}{42}(x - 1), \quad x \in [1;\, 2]; \end{aligned}$$

$$i=2:\ S_{II}(x)=q_1\frac{(x_2-x)^3}{6\cdot 1}+q_2\frac{(x-x_1)^3}{6\cdot 1}+$$
$$+\left(\frac{y_1}{1}-q_1\frac{1}{6}\right)(x_2-x)+\left(\frac{y_2}{1}-q_2\frac{1}{6}\right)(x-x_1)=\frac{18}{42}(3-x)^3+$$
$$+\left(-\frac{30}{42}\right)(x-2)^3+\frac{108}{42}(3-x)+\frac{282}{42}(x-2),\quad x\in[2;\,3];$$

$$i=3:\ S_{III}(x)=q_2\frac{(x_3-x)^3}{6\cdot 1}+q_3\frac{(x-x_2)^3}{6\cdot 1}+$$
$$+\left(\frac{y_2}{1}-q_2\frac{1}{6}\right)(x_3-x)+\left(\frac{y_3}{1}-q_3\frac{1}{6}\right)(x-x_2)=-\frac{30}{42}(4-x)^3+$$
$$+\frac{102}{42}(x-3)^3+\frac{282}{42}(4-x)+\frac{276}{42}(x-3),\quad x\in[3;\,4];$$

$$i=4:\ S_{IV}(x)=q_3\frac{(x_4-x)^3}{6\cdot 1}+q_4\frac{(x-x_3)^3}{6\cdot 1}+$$
$$+\left(\frac{y_3}{1}-q_3\frac{1}{6}\right)(x_4-x)+\left(\frac{y_4}{1}-q_4\frac{1}{6}\right)(x-x_3)=$$
$$=\frac{102}{42}(5-x)^3+\frac{276}{42}(5-x)+21(x-4),\quad x\in[4,\,5].$$

Проверим правильность построения сплайнов для узла $x^*=$ $=2$. К нему примыкают сплайны $S_{\mathrm{I}}(x)$ и $S_{\mathrm{II}}(x)$:

$$S_1(2-0)=3;\qquad S_{II}(2+0)=3;$$
$$S_1'(2-0)=\frac{120}{42};\qquad S_{II}'(2+0)=\frac{120}{42};$$
$$S_1''(2-0)=\frac{108}{42};\qquad S_{II}''(2+0)=\frac{108}{42}.$$

УПРАЖНЕНИЯ.

3.1 Выписать интерполяционные многочлены Лагранжа и Ньютона для узловых значений y_i, заданных функций $y=$ $=f(x)$ в точках x_i. Вычислить значение многочлена и оценить погрешность в точке x^*.

а) $y=\cos x$; $x_0=0$; $x_1=\pi/6$; $x_2=\pi/3$;
$x_3=\pi/2$; $x_4=2\pi/3$; $x^*=1$.

б) $y=\operatorname{tg} x$; $x_0=-\pi/3$; $x_1=-\pi/6$; $x_2=0$;
$x_3=\pi/6$; $x_4=\pi/3$; $x^*=0,5$.

в) $y = \ln x$; $x_0 = 1$; $x_1 = 1,5$; $x_2 = 2$;
$x_3 = 2,5$; $x_4 = 3,0$; $x^* = 2,3$.

г) $y = e^x$; $x_0 = 0$; $x_1 = 0,1$; $x_2 = 0,2$;
$x_3 = 0,3$; $x_4 = 0,4$; $x^* = 0,25$.

д) $y = e^{x^2}$; $x_0 = 1$; $x_1 = 1,1$; $x_2 = 1,2$;
$x_3 = 1,3$; $x_4 = 1,4$; $x^* = 1,25$.

е) $y = x^{-1}$; $x_0 = 1$; $x_1 = 2$; $x_2 = 3$; $x_3 = 4$; $x_4 = 5$; $x^* = 2,5$.

ж) $y = \sin x$; $x_0 = 0$; $x_1 = \pi/6$; $x_2 = \pi/3$;
$x_3 = \pi/2$; $x_4 = 2\pi/3$; $x^* = 1$.

3.2. В заданиях упражнения 3.1 оценить погрешность многочленной интерполяции для случая, когда полученные в них таблицы считаются экспериментальными (т. е. после формирования таблицы считать, что $f(x)$ отсутствует).

3.3. Выписать интерполяционные многочлены Ньютона, используя разделенные разности.

а) $y = \sin x$; $x_0 = 0$; $x_1 = \pi/6$; $x_2 = \pi/4$; $x_3 = \pi/2$.
б) $y = \cos x$; $x_0 = 0$; $x_1 = \pi/6$; $x_2 = \pi/4$; $x_3 = \pi/2$.
в) $y = e^{-x}$; $x_0 = 0$; $x_1 = 0,5$; $x_2 = 0,8$; $x_3 = 1,5$.
г) $y = \mathrm{tg}\, x$; $x_0 = 0$; $x_1 = \pi/6$; $x_2 = \pi/4$; $x_3 = \pi/3$.

3.4. Для заданий упражнения 3.1 выписать интерполяционные кубические сплайны дефекта один на каждом отрезке $x \in [x_{i-1}, x_i]$, $i = \overline{1,4}$. Проверить непрерывность $S(x_2)$, $S'(x_2)$, $S''(x_2)$.

§ 3.3. Метод наименьших квадратов

При наличии значительного числа экспериментальных точек сглаживание с помощью многочленной интерполяции не имеет смысла не только из-за неустойчивости (локальных выбросов) интерполирующей функции, но и из-за сильного колебания заданных точек. Способы локальной интерполяции, например с помощью сплайнов, также не дают приемлемых результатов.

В этом случае дискретно заданную функцию сглаживают в среднем, чаще всего многочленом, коэффициенты которого находят с помощью минимизации отклонения сглаживающей функции от заданных точек в *некотором среднеинтегральном смысле* (рис. 3.7).

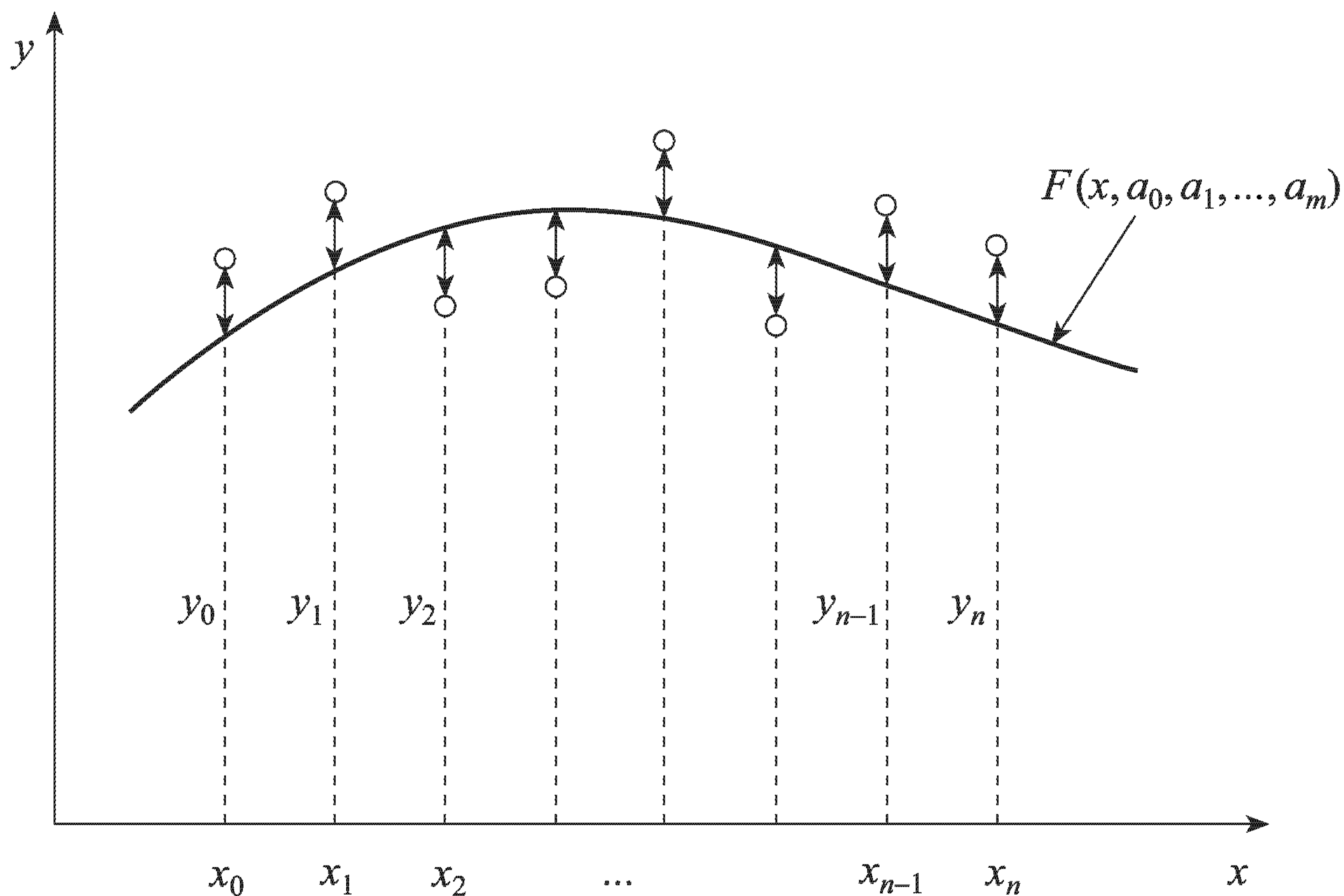


Рис. 3.7. К методу наименьших квадратов

Одним из таких методов является *метод наименьших квадратов* (МНК). Суть его заключается в следующем.

Пусть дана экспериментальная таблица (3.1). Составим многочлен степени m $(m \ll n)$

$$F(x, a_0, a_1, \ldots, a_m) = a_0 x^m + a_1 x^{m-1} + \ldots + a_m. \qquad (3.28)$$

Будем полагать, что отклонение значений этого многочлена от заданных таблицей (3.1) значений y_i, $i = \overline{0, n}$, минимально в некотором среднеинтегральном смысле.

В точечном методе наименьших квадратов строится функционал

$$S(a_0, a_1, \ldots, a_m) = \sum_{i=0}^{n} \left[F(x_i, a_0, a_1, \ldots, a_m) - y_i\right]^2, \qquad (3.29)$$

который геометрически представляет собой сумму квадратов отклонений значений y_i от значений аппроксимирующего многочлена (3.28) в точках x_i, $i = \overline{0, n}$ (на рис. 3.7 показаны двусторонними стрелками).

Необходимым условием минимума функции многих переменных является равенство нулю ее частных производных первого порядка по независимым переменным. В функционале (3.29) такими независимыми переменными являются коэффициенты $a_0, a_1, \ldots, a_m$ многочлена (3.28), которые до их определения являются не постоянными, а варьируемыми переменными:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2\sum\limits_{i=0}^{n}[F(x_i, a_0, a_1, \ldots, a_m) - y_i]\, x_i^m = 0, \\ \dfrac{\partial S}{\partial a_1} = 2\sum\limits_{i=0}^{n}[F(x_i, a_0, a_1, \ldots, a_m) - y_i]\, x_i^{m-1} = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial S}{\partial a_m} = 2\sum\limits_{i=0}^{n}[F(x_i, a_0, a_1, \ldots, a_m) - y_i]\, x_i^0 = 0. \end{cases} \tag{3.30}$$

Неоднородная СЛАУ (3.30) порядка $m + 1$ относительно неизвестных $a_0, a_1, \ldots, a_m$ является нормальной, и, следовательно, ее матрица является симметрической и положительно определенной. Решения $a_0, a_1, \ldots, a_m$ доставляют *минимум* функционалу (3.29).

Если представить СЛАУ (3.30) в виде

$$\begin{cases} b_{00}a_0 + b_{01}a_1 + \ldots + b_{0m}a_m = c_0, \\ b_{10}a_0 + b_{11}a_1 + \ldots + b_{1m}a_m = c_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_{m0}a_0 + b_{m1}a_1 + \ldots + b_{mm}a_m = c_m, \end{cases}$$

то можно выписать выражения для коэффициентов $b_{ij} = b_{ji}$, $i, j = \overline{0, m}$, и правых частей c_i, $i = \overline{0, m}$:

$$b_{00} = \sum_{i=0}^{n} x_i^{2m}, \quad b_{01} = \sum_{i=0}^{n} x_i^{2m-1}, \quad \ldots$$

$$\ldots, \quad b_{0m} = \sum_{i=0}^{n} x_i^{m}, \quad c_0 = \sum_{i=0}^{n} y_i x_i^{m};$$

$$b_{10} = \sum_{i=0}^{n} x_i^{2m-1}, \quad b_{11} = \sum_{i=0}^{n} x_i^{2m-2}, \quad \ldots$$

$$\ldots, \quad b_{1m} = \sum_{i=0}^{n} x_i^{m-1}, \quad c_1 = \sum_{i=0}^{n} y_i x_i^{m-1};$$

....

$$b_{m0} = \sum_{i=0}^{n} x_i^{m}, \quad b_{m1} = \sum_{i=0}^{n} x_i^{m-1}, \quad \ldots$$

$$\ldots, \quad b_{mm} = n+1, \quad c_m = \sum_{i=0}^{n} y_i.$$

Решив эту СЛАУ относительно a_0, a_1, ... , a_m и подставив их в (3.28), получаем многочлен степени $m \ll n$, наилучшим образом сглаживающий дискретную функцию (3.1) в среднеквадратическом смысле.

Для выбора степени многочлена (3.28) можно начинать с многочлена 1-й степени $F(x, a_0, a_1) = a_0 x + a_1$, решив задачу о точечном МНК для которого, находят коэффициенты a_0, a_1. Определяют максимальную по модулю величину относительной погрешности: $|\delta(y_i)|_m = \max\limits_i |(F(x_i) - y_i)/y_i|$, и если она не превышает заданную точность ε, то за аппроксимирующий многочлен принимают $F(x) = a_0 x + a_1$. В противном случае увеличивают степень многочлена на единицу и повторяют расчет и т. д.

Однако на практике заданную таблицу представляют графически и по заданным точкам приближенно определяют вид аппроксимирующего по МНК многочлена.

В интегральном методе наименьших квадратов рассматривается интегрируемая с квадратом функция $y = f(x)$, $x \in [a, b]$, которая трудна для исследования (например, трудно вычислить производные).

Будем аппроксимировать эту функцию некоторой функцией с минимизацией заштрихованной площади (см. рис. 3.8), например с помощью многочлена

$$F(x, a_0, a_1, \ldots, a_m) = a_0 x^m + a_1 x^{m-1} + \ldots + a_m,$$

где $a_0, a_1, \ldots, a_m$ находят из условия минимизации следующего квадратичного функционала:

$$S(a_0, a_1, \ldots, a_m) = \int_a^b [F(x, a_0, a_1, \ldots, a_m) - f(x)]^2\, dx. \quad (3.31)$$

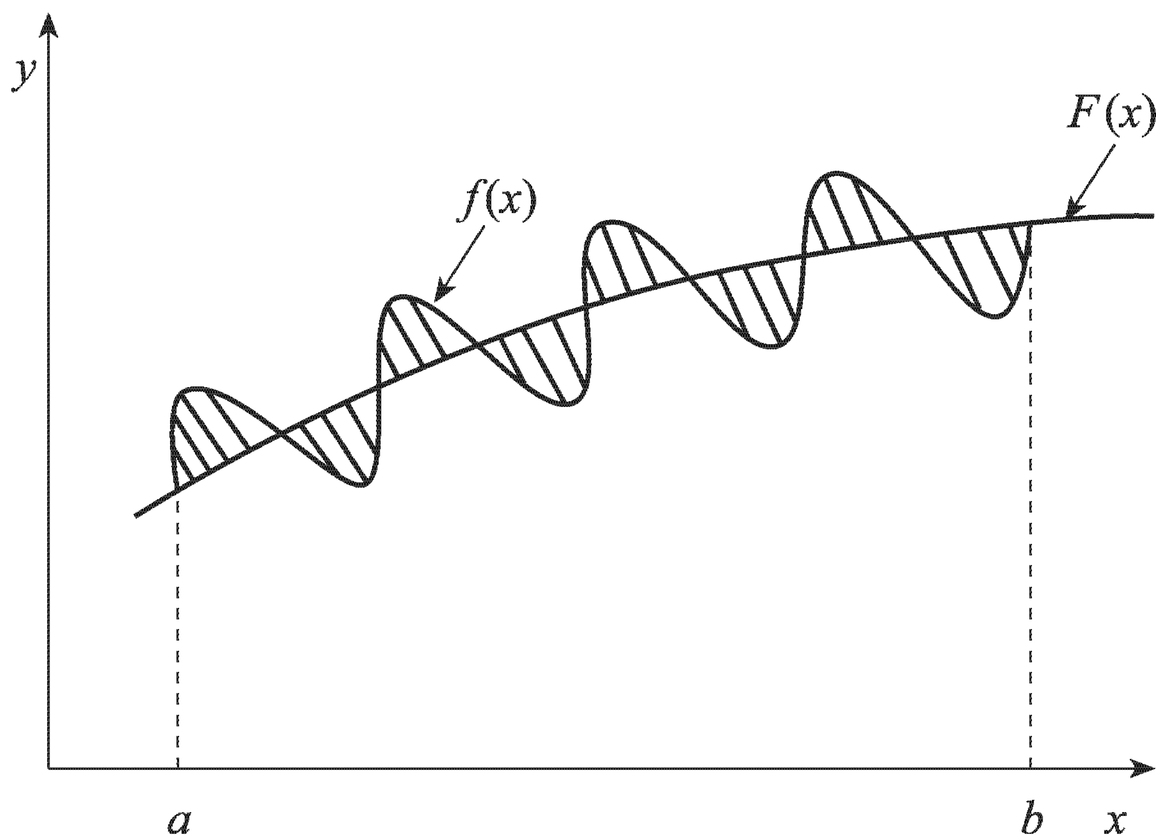


Рис. 3.8. К интегральному методу наименьших квадратов

Необходимые условия минимума функционала (3.31) имеют вид

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2\int\limits_a^b [F(x, a_0, \ldots, a_m) - f(x)]x^m dx = 0, \\ \dfrac{\partial S}{\partial a_1} = 2\int\limits_a^b [F(x, a_0, \ldots, a_m) - f(x)]x^{m-1} dx = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial S}{\partial a_m} = 2\int\limits_a^b [F(x, a_0, \ldots, a_m) - f(x)]x^0 dx = 0, \end{cases}$$

которые после интегрирования приобретают форму следующей СЛАУ:

$$\begin{cases} b_{00}a_0 + b_{01}a_1 + \ldots + b_{0m}a_m = \int\limits_a^b f(x)x^m dx, \\ b_{10}a_0 + b_{11}a_1 + \ldots + b_{1m}a_m = \int\limits_a^b f(x)x^{m-1} dx, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_{m0}a_0 + b_{m1}a_1 + \ldots + b_{mm}a_m = \int\limits_a^b f(x)x^0 dx. \end{cases} \quad (3.32)$$

В нормальной СЛАУ (3.32) относительно коэффициентов $a_0, a_1, \ldots, a_m$ правые части могут не интегрироваться в силу сложности исследуемой функции $f(x)$. В этом случае правые

части вычисляются с помощью методов численного интегрирования, которые рассматриваются ниже, в § 3.5.

Пример 3.5. Используя точечный метод наименьших квадратов аппроксимировать заданную таблицу линейным и квадратичным многочленами.

i	0	1	2	3	4	5
x_i	0	1	2	3	4	5
y_i	2,8	6,1	10,9	18,1	27,3	38

$F_1(x_i)$	−0,4285	6,6229	13,6743	20,7257	27,7771	34,8285
$F_2(x_i)$	2,82	5,9743	11,0796	18,1287	27,1287	38,078
$\delta_I(y_i)$	−1,153	0,086	0,255	0,145	0,0174	−0,083
$\delta_{II}(y_i)$	0,0071	−0,0206	0,0165	0,0016	−0,0063	0,0021

Решение. Под заданной таблицей необходимо заготовить таблицу для значений линейного $F_1(x_i)$ и квадратичного $F_2(x_i)$ многочленов, которые будут получены в результате решения задачи, а также для их относительных погрешностей.

Для линейного многочлена $(m = 1)$ $F_1(x, a_0, a_1) = a_0 x + a_1$ составляется функционал $(n = 5)$ $S(a_0, a_1) = \sum_{i=0}^{5} [(a_0 x_i + a_1) - y_i]^2$, частные производные от которого по неизвестным параметрам a_0, a_1 приравниваются нулю:

$$\frac{\partial S}{\partial a_0} = 2\sum_{i=0}^{5} [(a_0 x_i + a_1) - y_i]\, x_i = 0,$$

$$\frac{\partial S}{\partial a_1} = 2\sum_{i=0}^{5} [(a_0 x_i + a_1) - y_i] \cdot 1 = 0.$$

В результате получается СЛАУ 2-го порядка относительно неизвестных параметров a_0, a_1:

$$\begin{cases} b_{00} a_0 + b_{01} a_1 = c_0, \\ b_{10} a_0 + b_{11} a_1 = c_1, \end{cases}$$

где $b_{00} = \sum\limits_{i=0}^{5} x_i^2$; $b_{01} = \sum\limits_{i=0}^{5} x_i$; $c_0 = \sum\limits_{i=0}^{5} y_i x_i$; $b_{10} = \sum\limits_{i=0}^{5} x_i$; $b_{11} = \sum\limits_{i=0}^{5} x_i^0 = 6$; $c_1 = \sum\limits_{i=0}^{5} y_i$.

Используя табличные значения x_i, y_i $i = \overline{0, 5}$, получим СЛАУ

$$\begin{cases} 55a_0 + 15a_1 = 381{,}4, \\ 15a_0 + 6a_1 = 103{,}2. \end{cases}$$

Решая ее с помощью правила Крамера:

$$a_0 = \frac{\Delta a_0}{\Delta}; \quad a_1 = \frac{\Delta a_1}{\Delta}; \quad \Delta = \begin{vmatrix} 55 & 15 \\ 15 & 6 \end{vmatrix} = 105 \neq 0;$$

$$\Delta a_0 = \begin{vmatrix} 381{,}4 & 15 \\ 103{,}2 & 6 \end{vmatrix} = 740{,}4; \quad \Delta a_1 = \begin{vmatrix} 55 & 381{,}4 \\ 15 & 103{,}2 \end{vmatrix} = -45,$$

находим $a_0 = 7{,}0514$; $a_1 = -0{,}4285$; $F_1(x) = 7{,}0514x - 0{,}4285$. Значения этого многочлена в заданных узлах x_i, $i = \overline{0;\, 5}$, заносятся в первую строку дополнительной таблицы, а относительные погрешности $\delta_{\mathrm{I}}(y_i) = (F_1(x_i) - y_i)/y_i$ — в третью строку дополнительной таблицы. Максимальная по модулю относительная погрешность $\max\limits_i |\delta_{\mathrm{I}}(y_i)| = 1{,}153$, т. е. более 100%, что значительно превышает заданную точность.

Поэтому далее рассматривается квадратичная аппроксимирующая функция $F_2(x, a_0, a_1, a_2) = a_0 x^2 + a_1 x + a_2$, для которой функционал и нормальная СЛАУ имеют вид

$$S(a_0, a_1, a_2) = \sum_{i=0}^{5} \left[\left(a_0 x_i^2 + a_1 x_i + a_2 \right) - y_i \right]^2,$$

$$\frac{\partial S}{\partial a_0} = 2 \sum_{i=0}^{5} \left[\left(a_0 x_i^2 + a_1 x_i + a_2 \right) - y_i \right] x_i^2 = 0,$$

$$\frac{\partial S}{\partial a_1} = 2 \sum_{i=0}^{5} \left[\left(a_0 x_i^2 + a_1 x_i + a_2 \right) - y_i \right] x_i = 0,$$

$$\frac{\partial S}{\partial a_2} = 2 \sum_{i=0}^{5} \left[\left(a_0 x_i^2 + a_1 x_i + a_2 \right) - y_i \right] \cdot 1 = 0.$$

Собирая коэффициенты при неизвестных a_0, a_1, a_2, получаем

$$\begin{cases} b_{00}a_0 + b_{01}a_1 + b_{02}a_2 = c_0, \\ b_{10}a_0 + b_{11}a_1 + b_{12}a_2 = c_1, \\ b_{20}a_0 + b_{21}a_1 + b_{22}a_2 = c_2, \end{cases}$$

где $b_{00} = \sum\limits_{i=0}^{5} x_i^4$; $b_{01} = b_{10} = \sum\limits_{i=0}^{5} x_i^3$; $b_{02} = b_{20} = \sum\limits_{i=0}^{5} x_i^2$; $b_{21} = b_{12} = \sum\limits_{i=0}^{5} x_i$; $b_{11} = \sum\limits_{i=0}^{5} x_i^2$; $b_{22} = 6$. $c_0 = \sum\limits_{i=0}^{5} y_i x_i^2$; $c_1 = \sum\limits_{i=0}^{5} y_i x_i$; $c_2 = \sum\limits_{i=0}^{5} y_i$.

Тогда в соответствии с табличными значениями x_i, y_i имеем (каждое уравнение разделено на коэффициент при a_0)

$$\begin{cases} a_0 + 0{,}2298a_1 + 0{,}05618a_2 = 1{,}6337, \\ a_0 + 0{,}2444a_1 + 0{,}06667a_2 = 1{,}6951, \\ a_0 + 0{,}2727a_1 + 0{,}1091a_2 = 1{,}8764, \end{cases}$$

откуда получаем $a_0 = 0{,}9743$; $a_1 = 2{,}18$; $a_2 = 2{,}82$. Таким образом,

$$F_2(x) = 0{,}9743x^2 + 2{,}18x + 2{,}82.$$

Значения этого многочлена заносятся во вторую строку дополнительной таблицы, а относительные погрешности $\delta_{\text{II}}(y_i) = (F_2(x_i) - y_i)/y_i$ — в четвертую строку этой таблицы, откуда видно, что максимальная по модулю погрешность $\max\limits_i |\delta_{\text{II}}(y_i)| = 0{,}0206$, что меньше заданной точности $\varepsilon = 0{,}05$.

Таким образом, многочлен $F_2(x)$ наилучшим образом в квадратичном смысле приближает заданную таблицу с точностью $\varepsilon = 0{,}05$.

УПРАЖНЕНИЯ.

3.5. Точечным методом наименьших квадратов аппроксимировать линейным и квадратичным многочленами следующие дискретно заданные функции и определить максимальную по модулю погрешность аппроксимации:

а)

x_i	0	1	2	3	4	5
y_i	4,2	8,8	16,3	24,6	36,5	48,4

б)

x_i	0	1	2	3	4	5
y_i	3,2	7,8	15,3	23,6	35,5	47,5

в)

x_i	0	1	2	3	4	5
y_i	1,9	5,2	9,8	17,3	25,7	37,5

г)

x_i	0	1	2	3	4	5
y_i	1,1	3,9	9,2	16,8	25,3	35,7

д)

x_i	0	1	2	3	4	5
y_i	1,1	4,9	11,2	18,8	29,3	40,7

е)

x_i	0	1	2	3	4	5
y_i	1,1	5,9	13,2	21,8	33,4	45,4

ж)

x_i	0	1	2	3	4	5
y_i	2,1	6,9	13,7	23,4	33,6	47,5

з)

x_i	0	1	2	3	4	5
y_i	3,1	5,8	11,2	17,7	27,4	37,5

3.6. Показать, что в методе наименьших квадратов решения $a_0, a_1, \ldots, a_m$ СЛАУ (3.30) доставляют минимум (а не максимум) функционалу (3.29).

3.7. Интегральным методом наименьших квадратов аппроксимировать линейным, квадратичным и кубичным многочленами следующие функции, заданные на отрезках $x \in [a;\ b]$, и определить относительную погрешность:

а) $y = \sin(x)$, $x \in [-\pi/4;\ \pi/4]$;
б) $y = \operatorname{tg} x$, $x \in [-\pi/4;\ \pi/4]$;
в) $y = e^x$, $x \in [-1;\ 1]$;
г) $y = e^{-x}$, $x \in [-1;\ 1]$;
д) $y = \ln x$, $x \in [0,5;\ 2]$;
е) $y = \cos(x)$, $x \in [-\pi/4;\ \pi/4]$.

§ 3.4. Численное дифференцирование

Этот параграф является основополагающим для численного интегрирования дифференциальных уравнений, как обыкновенных, так и в частных производных.

Пусть в некоторой точке x^* требуется вычислить производные первого, второго и т. д. порядков от дискретно заданной функции (3.1). Могут иметь место следующие два случая: а) точка $x^* \in (x_{i-1}, x_i)$, $i = \overline{1, n}$, и б) точка $x^* = x_i$, $i = \overline{1, n-1}$, т. е. совпадает с одним из внутренних узлов заданной таблицы.

Тогда *в первом случае* заданная таблица сглаживается какой-либо функцией $\varphi\,(x)$, являющейся глобальным (локальным) интерполяционным многочленом или многочленом, полученным по МНК с некоторой погрешностью $R_n(x)$, в результате чего имеют место следующие равенства

$$\begin{aligned}
y(x) &= \varphi(x) + R_n(x), & y(x^*) &= \varphi(x^*) + R_n(x^*);\\
y'(x) &= \varphi'(x) + R'_n(x), & y'(x^*) &= \varphi'(x^*) + R'_n(x^*);\\
y''(x) &= \varphi''(x) + R''_n(x), & y''(x^*) &= \varphi''(x^*) + R''_n(x^*);\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{3.33}$$

Следует отметить, что процедура численного дифференцирования является *некорректной* в том смысле, что близость искомой функции $y(x)$ и сглаживающей функции $\varphi\,(x)$ не гарантирует близости их производных (рис. 3.9). Более того, они могут иметь в одной и той же точке x^* производные различных знаков. Тем не менее формулы (3.33) широко используются на практике.

Во *втором случае* ($x^* = x_i$, $i = \overline{1, n-1}$) используется аппарат разложения функций в ряд Тейлора, для чего функция в точке x^* должна иметь достаточное число производных. С этой целью предполагается, что заданная таблица (3.1) является сеточной функцией для некоторой функции $y(x)$, имеющей в точке x^* производные до четвертого порядка включительно, т. е. что $y_i = y(x_i)$.

Тогда внутренний узел $x^* = x_i$, $i = \overline{1, n-1}$, окружают узлы x_{i-1}, x_{i+1} (рис. 3.10), причем $x_{i+1} - x_i = x_i - x_{i-1} = h$. Тогда, разлагая значения y_{i-1}, y_{i+1} *на точной функции* в ряд Тейлора в окрестности точки x_i до производной четвертого по-

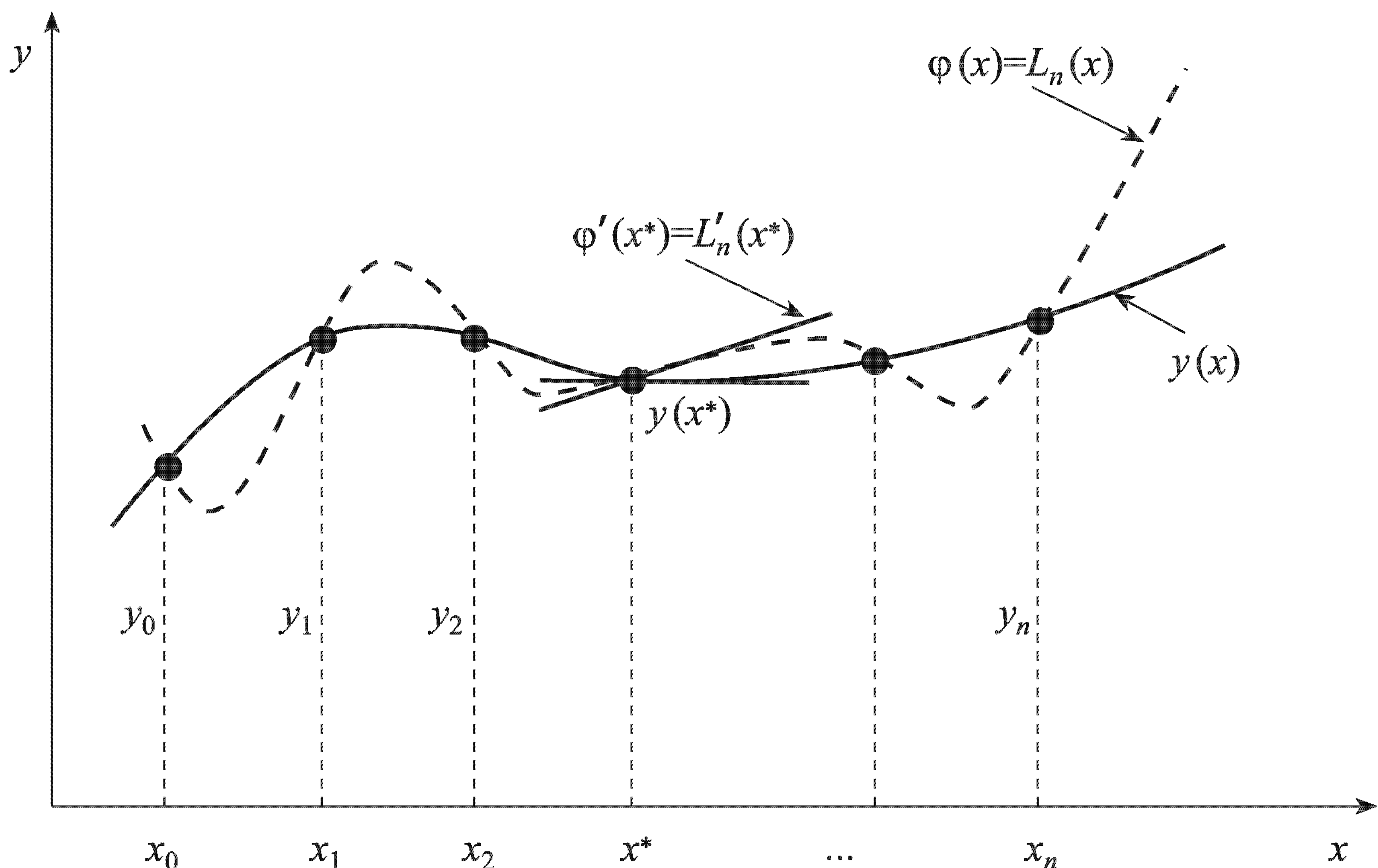


Рис. 3.9. К понятию некорректности численного дифференцирования

рядка включительно, получим

$$y_{i-1} = y(x_i - h) = y_i - y_i' h + y_i'' \frac{h^2}{2} - \\ - y_i''' \frac{h^3}{6} + y^{IV}(\xi)\frac{h^4}{24}, \quad \xi \in (x_{i-1}, x_i), \quad (3.34)$$

$$y_{i+1} = y(x_i + h) = y_i + y_i' h + y_i'' \frac{h^2}{2} + \\ + y_i''' \frac{h^3}{6} + y^{IV}(\xi)\frac{h^4}{24}, \quad \xi \in (x_i, x_{i+1}). \quad (3.35)$$

Выразим вначале y_i' из (3.34), а затем из (3.35), разделив предварительно на h и оставляя слагаемые с первой степенью шага h, получим

$$\overline{y}_i' = \frac{y_i - y_{i-1}}{h} + O(h) = \frac{\Delta \overline{y}_i}{h} + O(h), \quad (3.36)$$

$$y_i' = \frac{y_{i+1} - y_i}{h} + O(h) = \frac{\Delta y_i}{h} + O(h). \quad (3.37)$$

Вычтем из (3.35) выражение (3.34), разделим полученное соотношение на $2h$, получим следующее значение производной первого порядка в точке x_i (слагаемые с производными четного

порядка сокращаются за исключением слагаемого с производной четвертого порядка):

$$\overset{\circ}{y}{}'_i = \frac{y_{i+1} - y_{i-1}}{2h} + O(h^2) = \frac{\Delta \overset{\circ}{y}_i}{2h} + O(h^2), \qquad (3.38)$$

где $\Delta \overset{\circ}{y}_i = y_{i+1} - y_{i-1}$ — центральная разность первого порядка.

Выражение (3.36) определяет производную первого порядка в узле x_i с помощью отношения конечных разностей *слева*. Она имеет первый порядок аппроксимации относительно шага h.

Выражение (3.37) определяет производную первого порядка в узле x_i с помощью отношения конечных разностей *справа*. Она также имеет первый порядок аппроксимации относительно шага h.

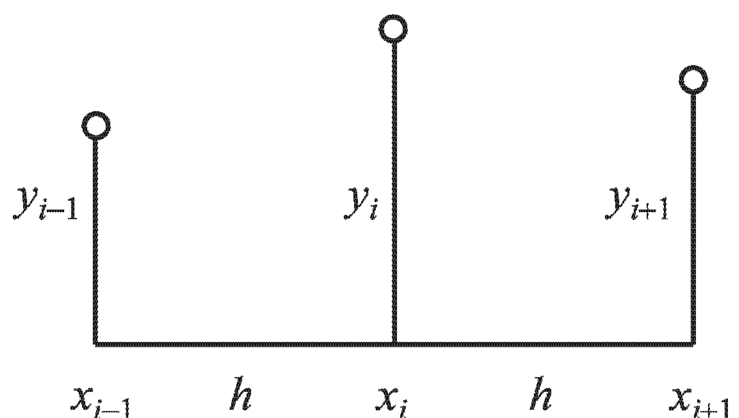


Рис. 3.10. Численное дифференцирование с помощью отношения конечных разностей

Выражение (3.38) определяет производную первого порядка в узле x_i с помощью отношения *центральных разностей*. Она имеет второй порядок аппроксимации относительно шага h.

Сложим выражения (3.34) и (3.35), разделим на h^2 (слагаемые с производными нечетного порядка сокращаются), получим

$$y''(x_i) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + O(h^2) = \frac{\Delta^2 y_i}{h^2} + O(h^2). \qquad (3.39)$$

Выражение (3.39) определяет производную второго порядка в узле x_i с помощью отношения центральных разностей второго порядка. Она имеет второй порядок аппроксимации относительно шага h.

Определим *порядок точности метода* численного дифференцирования с помощью отношения конечных разностей, для чего рассмотрим, как из выражений (3.34), (3.35) получаются производные (3.36)–(3.38). Например, производная (3.36) получена из (3.34) следующим образом:

$$\overline{y}'_i = \frac{y_i - y_{i-1}}{h} + y''_i \frac{h}{2} - y'''_i \frac{h^3}{2} + O(h^3).$$

Поскольку шаг h является дробной величиной, то главный член погрешности содержит первую степень h $\left(\text{слагаемое } y_i'' \frac{h}{2}\right)$, остальные члены погрешности являются членами более высокого порядка малости (слагаемые, начиная с h^2 и далее), поэтому, отбрасывая их и сохраняя главный член погрешности, следующий сразу за отношением конечных разностей, получим формулу (3.36). Аналогично получены и остальные формулы.

Порядком точности метода численного дифференцирования с помощью отношения конечных разностей называют показатель степени h в главном члене погрешности. (По поводу связи погрешности и точности см. гл. I и теорему эквивалентности в п. 6.3.4).

Таким образом, односторонние производные (3.36) и (3.37) имеют 1-й порядок точности и поэтому менее точны, чем центрально-разностная производная (3.38), имеющая второй порядок точности.

3.4.1. Метод Рунге уточнения формул численного дифференцирования. Из формул (3.36)–(3.38) видно, что метод p-го порядка численного дифференцирования совпадает с показателем степени шага h в главном члене погрешности и имеет вид

$$f'(x) = \varphi_h'(x) + h^p \cdot \psi(x) + O\left(h^{p+1}\right) + O\left(h^{p+2}\right) + \ldots, \qquad (3.40)$$

где

$$f'(x) \approx \varphi_h'(x),$$

а остаточный член имеет вид

$$R_p = h^p \psi(x) + O\left(h^{p+1}\right) + O\left(h^{p+2}\right) + \ldots$$

С целью повышения на единицу порядка точности метода продифференцируем численно методом p-го порядка функцию $f(x_i) = y_i$, $i = \overline{0, n}$, с шагом h, получим выражение (3.40). Затем продифференцируем численно функцию тем же методом p-го порядка, с шагом kh, ($k = 1/2$; $1/4$; $1/16$; ...), получим

$$f'(x) = \varphi_{kh}'(x) + (kh)^p \psi(x) + O\left(h^{p+1}\right). \qquad (3.41)$$

Вычитая из (3.41) выражение (3.40) и определяя из полученного равенства $\psi(x)$, находим

$$h^p\psi(x) = \frac{\varphi'_h(x) - \varphi'_{kh}(x)}{k^p - 1} + O\left(h^{p+1}\right). \qquad (3.42)$$

Выражение (3.42) можно использовать для оценки погрешности численного дифференцирования.

Подставляя (3.42) в (3.41), получаем окончательно

$$f'(x) = \varphi'_{kh}(x) + k^p\frac{\varphi'_h(x) - \varphi'_{kh}(x)}{k^p - 1} + O\left(h^{p+1}\right). \qquad (3.43)$$

Из (3.43) видно, что это уже метод порядка $p+1$, т. е. на порядок точнее. В этом и заключается процедура Рунге уточнения численного дифференцирования.

Пример 3.6. Вычислить производную в точке $x^* = 0,5$ от дискретно заданной функции примера 3.1. Оценить погрешность производной.

Решение. По заданной таблице составляем интерполяционный многочлен

$$L_2(x) = \frac{2}{3}x^2 + \frac{4}{3}x + 1$$

(см. пример 3.1) и погрешность

$$R_2(x) = \frac{3{,}978}{6}(x - x_0)(x - x_1)(x - x_2) = \\ = \frac{3{,}978}{6}(x+1)\cdot x\cdot(x-1).$$

Тогда

$$y(x) = L_2(x) + R_2(x);$$

$$y'(x) = L'_2(x) + R'_2(x);$$

$$y'(x^*) = L'_2(x^*) + R'_2(x^*) = \left(\frac{2}{3}x^2 + \frac{4}{3}x + 1\right)'_{x=x*} + \\ + \frac{3{,}978}{6}\left[(x+1)\,x\,(x-1)\right]'_{x=x*} = \left(\frac{4}{3}x^* + \frac{4}{3}\right) + \frac{3{,}978}{6}\left(3x^{*2} - 1\right).$$

$$y'(0{,}5) = \left(\frac{4}{3}\cdot 0{,}5 + \frac{4}{3}\right) + \frac{3{,}978}{6}\left(3\cdot 0{,}5^2 - 1\right) = 2 - 0{,}166.$$

То есть

$$y'(0{,}5) \approx 2; \quad R'_2(0{,}5) \approx -0{,}166.$$

Точное значение производной первого порядка в точке $x^* = 0{,}5$ будет

$$y' = (3^x)'_{x=x^*} = 3^{x^*} \cdot \ln 3 = 1{,}9028.$$

Аналогично вычисляются и производные 2-го, 3-го и т. д. порядков. В данном примере производная второго порядка постоянна для всего отрезка $[-1;\, 1]$ и равна $y'' \approx 4/3$.

Пример 3.7. Для узла x_i заданной таблицы с помощью отношения конечных разностей слева, справа и отношения центральных разностей выписать выражения для производной третьего порядка. Определить порядок аппроксимации во всех трех случаях.

Решение. Заметим, что из формул численного дифференцирования минимальное количество узлов, необходимое для вычисления конечных разностей какого-либо порядка, должно быть на единицу больше этого порядка. Следовательно, для вычисления конечных разностей третьего порядка, входящих в конечно-разностную производную, необходимо не менее четырех узлов.

В соответствии с этим рассмотрим таблицу с пятью узлами, отстоящими друг от друга на одинаковом расстоянии $h = x_k - x_{k-1}$, $k = i-1,\, i,\, i+1,\, i+2$.

x_{i-2}	x_{i-1}	x_i	x_{i+1}	x_{i+2}
y_{i-2}	y_{i-1}	y_i	y_{i+1}	y_{i+2}

Тогда для вычисления производных 3-го порядка в узле x_i разложим y_{i-2}, y_{i-1}, y_{i+1}, y_{i+2}, как значения дифференцируемой необходимое число раз функции $y = f(x)$, в ряд Тейлора в окрестности узла x_i до пятой производной включительно, получим

$$y_{i-2} = y(x_i - 2h) = \\ = y_i - y_i'(2h) + y_i''\frac{(2h)^2}{2} - y_i'''\frac{(2h)^3}{6} + y_i^{IV}\frac{(2h)^4}{24} + O\left(h^5\right),$$

$$y_{i-1} = y(x_i - h) = y_i - y_i'h + y_i''\frac{h^2}{2} - y_i'''\frac{h^3}{6} + y_i^{IV}\frac{h^4}{24} + O\left(h^5\right),$$

$$y_{i+1} = y(x_i + h) = y_i + y_i'h + y_i''\frac{h^2}{2} + y_i'''\frac{h^3}{6} + y_i^{IV}\frac{h^4}{24} + O\left(h^5\right),$$

$$y_{i+2} = y(x_i + 2h) = \\ = y_i + y_i'(2h) + y_i'' \frac{(2h)^2}{2} + y_i''' \frac{(2h)^3}{6} + y_i^{IV} \frac{(2h)^4}{24} + O\left(h^5\right).$$

Выпишем различные выражения для производной 3-го порядка с помощью отношений конечных разностей:

$$\overline{y}_i''' \approx \frac{\Delta^3 \overline{y}_i}{h^3} = \frac{y_{i+1} - 3y_i + 3y_{i-1} - y_{i-2}}{h^3},$$

$$y_i''' \approx \frac{\Delta^3 y_i}{h^3} = \frac{y_{i+2} - 3y_{i+1} + 3y_i - y_{i-1}}{h^3},$$

$$\overset{\circ}{y}{}_i''' = \frac{1}{2}\left(\overline{y}_i''' + y_i'''\right) = \frac{y_{i+2} - 2y_{i+1} + 2y_{i-1} - y_{i-2}}{2h^3}.$$

Подставляя в эти формулы выше приведенные разложения в ряды Тейлора, получим

$$\overline{y}_i''' = \frac{y_{i+1} - 3y_i + 3y_{i-1} - y_{i-2}}{h^3} + O(h),$$

$$y_i''' = \frac{y_{i+2} - 3y_{i+1} + 3y_i - y_{i-1}}{h^3} + O(h),$$

$$\overset{\circ}{y}{}_i''' = \frac{1}{2}\left(\overline{y}_i''' + y_i'''\right) + O\left(h^2\right),$$

откуда видно, что производные 3-го порядка в узле x_i, вычисленные с помощью отношения конечных разностей слева и справа, имеют первый порядок аппроксимации, а с помощью отношения центральных разностей — второй порядок.

УПРАЖНЕНИЯ.

3.8. В заданиях упражнения 3.5 вычислить в точке $x = 2,5$ значения производных первого и второго порядков. Оценить их погрешность.

3.9. В заданиях упражнения 3.5 с помощью отношения конечных разностей вычислить в узле $x_2 = 2$ производные до четвертого порядка включительно в виде отношения конечных разностей слева, справа и отношения центральных разностей. Определить порядок аппроксимации полученных выражений.

Указание: для производной 4-го порядка использовать следующее конечно-разностное соотношение:

$$y_i^{IV} \approx \Delta^4 y_i / h^4 = \left(y_{i+2} - 4y_{i+1} + 6y_i - 4y_{i-1} + y_{i-2}\right) / h^4.$$

3.10. В заданиях упражнения 3.5 в узле $x_2 = 2$ вычислить значения лево- и правосторонней производных первого порядка с помощью отношения конечных разностей с шагом $h = 1$ и $h = 0{,}5$ и уточнить эти производные с помощью процедуры Рунге.

§ 3.5. Численное интегрирование функций

Известно, что для подавляющего большинства функций не удается вычислить первообразные, вследствие чего приходится прибегать к методам приближенного и численного интегрирования функций. Методы приближенного интегрирования используют разложение подынтегральных функций в ряды Тейлора (Маклорена) и дальнейшего почленного интегрирования членов ряда. К недостаткам методов приближенного интегрирования относится требование дифференцируемости подынтегральных функций до порядка, который требуется при разложении функций в ряд Тейлора. От этого недостатка свободны методы численного интегрирования, в которых подынтегральная функция удовлетворяет только условию непрерывности (для существования определенного интеграла).

При численном интегрировании по заданной подынтегральной функции строится сеточная функция (3.1), затем эта функция с помощью формул локального интерполирования с контролируемой погрешностью заменяется интерполяционным многочленом, интеграл от которого хорошо вычисляется и сравнительно легко оценивается погрешность.

Пусть на отрезке $x \in [a, b]$ дана непрерывная функция $y = f(x)$, требуется на $x \in [a, b]$ вычислить определенный интеграл

$$I = \int_a^b f(x)\,dx. \tag{3.44}$$

Заменим данную функцию $f(x)$ на сеточную функцию (3.1). Вместо точного значения интеграла (3.44) будем искать его приближенное значение с помощью суммы:

$$I \approx I_h = \sum_{i=0}^{n} A_i h_i, \quad h_i = x_i - x_{i-1}, \;\; i = \overline{1, n}, \;\; x_0 = a, \;\; x_n = b, \tag{3.45}$$

в которой необходимо определить коэффициенты A_i и погрешность формулы (3.45).

3.5.1. Формула прямоугольников численного интегрирования. Наиболее простой (и имеющей малую точность) является формула прямоугольников. Она основана на определении определенного интеграла как предела последовательности интегральных сумм:

$$\int_a^b f(x)dx = \lim_{\substack{\max \\ \Delta x_i \to 0}} \sum_{i=1}^{n} f(\xi_i)\Delta x_i,\ \ \xi_i \in [x_i, x_{i-1}],\ \ \Delta x_i = x_i - x_{i-1}.$$

Если в этом определении снять знак предела и положить $\Delta x_i = h_i$, $i = \overline{1, n}$, то появится погрешность $R_{\text{пр}}$ (за ξ_i можно принять левый или правый конец отрезка Δx_i), т. е.

$$\int_a^b f(x)dx = \sum_{i=1}^{n} f(x_{i-1})h_i + R_{np}, \tag{3.46}$$

или

$$\int_a^b f(x)dx = \sum_{i=1}^{n} f(x_i)h_i + R_{np}. \tag{3.47}$$

Формулы (3.46), (3.47) — *формулы прямоугольников численного интегрирования.*

Рассмотрим погрешность R_i формулы прямоугольников (3.46) на одном шаге $[x_{i-1}, x_i]$ численного интегрирования.

Для этого предположим, что первообразная $F(x)$ для подынтегральной функции $f(x)$ (она существует, поскольку $f(x)$ непрерывна на отрезке $x \in [a, b]$) непрерывно дифференцируема. Тогда, разлагая $F(x_i)$ в окрестности узла x_{i-1} в ряд Тейлора до второй производной включительно и используя равенство $F'(x) = f(x) = y(x)$, получим

$$\begin{aligned} R_i = \int_{x_{i-1}}^{x_i} f(x)dx - y_{i-1}h &= [F(x_i) - F(x_{i-1})] - y_{i-1}h = \\ &= [F'(x_{i-1})h + F''(\xi)\frac{h^2}{2}] - y_{i-1}h = \\ &= [y_{i-1}h + y'(\xi)\frac{h^2}{2}] - y_{i-1}h = y'(\xi)\frac{h^2}{2}, \quad \xi \in (x_{i-1}, x_i). \end{aligned}$$

На всем отрезке $[a, b]$ эту погрешность необходимо просуммировать n раз $(b - a = nh)$, получим

$$R_{np} = R_i n = y'(\xi)\frac{(b-a)h}{2}, \quad \xi \in (a, b). \tag{3.48}$$

Поскольку местоположение точки ξ на интервале $x \in [a, b]$ неизвестно, то на основе погрешности (3.48) можно выписать верхнюю оценку абсолютной погрешности метода прямоугольников и при заданной точности ε метода выписать неравенства

$$|R_{\text{пр}}| \leqslant \frac{(b-a)h}{2} M_1 \leqslant \varepsilon, \quad M_1 = \max_{x\in[a,b]} \left|f'(x)\right|.$$

Последнее неравенство можно использовать для верхней оценки шага h численного интегрирования по методу прямоугольников:

$$h \leqslant \frac{2\varepsilon}{(b-a)\,M_1}, \quad M_1 = \max_{x\in[a,b]} \left|f'(x)\right|.$$

На всем отрезке $x \in [a, b]$ формула прямоугольников (3.46) имеет вид

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} y_{i-1} h_i. \tag{3.49}$$

Ее погрешность определяется выражением (3.48).

Таким образом, определяющими формулами метода прямоугольников являются формула (3.49) численного интегрирования и формула (3.48) погрешности.

Из (3.48) видно, что на каждом отрезке $[x_{i-1}, x_i]$ формула прямоугольников имеет погрешность, пропорциональную h^2, а на всем отрезке $x \in [a, b]$ — шагу численного интегрирования h. В соответствии с этим *метод прямоугольников является методом первого порядка точности* (*главный член погрешности пропорционален шагу в первой степени*).

3.5.2. Численное интегрирование с помощью формулы трапеций. Рассмотрим интеграл (3.44) на отрезке $x \in \in [x_{i-1}, x_i]$ и будем на этом отрезке вычислять его приближенно, заменяя подынтегральную функцию интерполяционным много-

членом Лагранжа первой степени, получим

$$\int_{x_{i-1}}^{x_i} f(x)dx = \int_{x_{i-1}}^{x_i} L_1(x)dx + R_i, \qquad (3.50)$$

где R_i — погрешность, которая подлежит определению (на рис. 3.11 заштрихована), а L_1 — интерполяционный многочлен

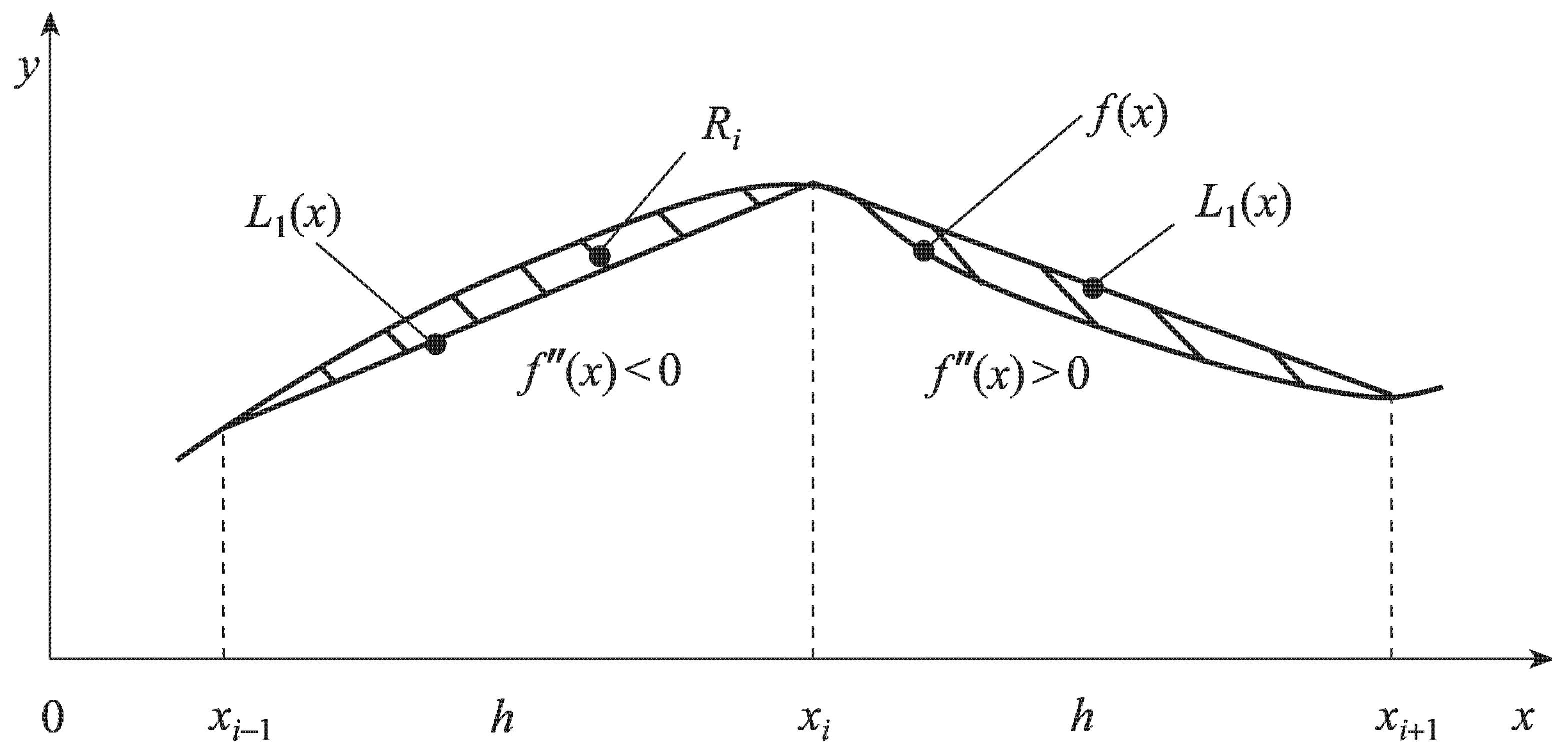


Рис. 3.11. К выводу формулы трапеций

Лагранжа первой степени, проведенный через два узла интерполяции x_{i-1} и x_i:

$$L_1(x) = y_{i-1}\frac{x - x_i}{x_{i-1} - x_i} + y_i\frac{x - x_{i-1}}{x_i - x_{i-1}}.$$

Пусть $x_i - x_{i-1} = h = \text{const}$, где $i = \overline{1, n}$. Обозначим

$$\frac{x - x_{i-1}}{h} = t,$$

тогда

$$\frac{x - x_i}{h} = \frac{(x - x_{i-1}) - (x_i - x_{i-1})}{h} = t - 1, \quad dx = hdt,$$

и многочлен Лагранжа примет вид

$$L_1(x) = L_1(x_{i-1} + ht) = -y_{i-1}(t - 1) + y_i t.$$

При $x = x_i$ верхний предел $t = 1$, при $x = x_{i-1}$ нижний предел $t = 0$.

Теперь интеграл в (3.50) от многочлена $L_1(x)$ можно представить в виде

$$\int\limits_{x_{i-1}}^{x_i} f(x)dx \approx \int\limits_{x_{i-1}}^{x_i} L_1(x)dx = h\int\limits_0^1 [-y_{i-1}(t-1)+y_i t]dt =$$
$$= h\left[-y_{i-1}(\frac{t^2}{2}-t)+y_i\frac{t^2}{2}\right]_0^1 = \frac{h}{2}(y_{i-1}+y_i). \quad (3.51)$$

Выражение (3.51) называют формулой трапеций численного интегрирования на отрезке $x \in [x_{i-1}, x_i]$.

Для всего отрезка $[a,\, b]$ необходимо сложить выражение (3.51) n раз:

$$\int\limits_a^b f(x)\,dx \approx \frac{h}{2}\left(y_0 + y_n + 2\sum_{i=1}^{n-1} y_i\right). \quad (3.52)$$

Выражение (3.52) *называют формулой трапеций численного интегрирования для всего отрезка* $[a,\, b]$.

Перейдем к оценке погрешности R_i формулы трапеций на отрезке $[x_{i-1}, x_i]$. Для этого будем предполагать, что подынтегральная функция принадлежит классу C^2 (дважды непрерывно дифференцируема), тогда для первообразной $F(x)$ существует производная 3-го порядка.

В этом случае погрешность на отрезке $[x_{i-1}, x_i]$ определяется следующим образом:

$$R_i = \int\limits_{x_{i-1}}^{x_i} f(x)dx - \frac{h}{2}(y_{i-1}+y_i) =$$
$$= [F(x_i) - F(x_{i-1})] - \frac{h}{2}(y_{i-1}+y_i). \quad (3.53)$$

Разложим $F(x_i)$ в окрестности точки x_{i-1} в ряд Тейлора до 3-й производной включительно, получим $(F(x_i) = F(x_{i-1}+h))$

$$F(x_i) - F(x_{i-1}) = -F(x_{i-1}) + F(x_{i-1}) +$$
$$+ F'(x_{i-1})h + F''(x_{i-1})\frac{h^2}{2} + F'''(\xi)\frac{h^3}{6}, \quad \xi \in (x_{i-1}, x_i);$$

$$F'(x_{i-1}) = f(x_{i-1}) = y_{i-1}, \quad F''(x_{i-1}) = f'(x_{i-1}) = y'_{i-1},$$
$$F'''(\xi) = f''(\xi) = y''(\xi).$$

Тогда

$$\int\limits_{x_{i-1}}^{x_i} f(x)dx = y_{i-1}h + y'_{i-1}\frac{h^2}{2} + y''(\xi)\frac{h^3}{6}, \quad \xi \in (x_{i-1}, x_i). \tag{3.54}$$

То же самое сделаем и со вторым слагаемым в левой части равенства (3.53), т. е. разложим y_i в окрестности точки x_{i-1} в ряд Тейлора до 2-ой производной включительно, получим

$$\frac{h}{2}(y_{i-1} + y_i) = \frac{h}{2}\left[y_{i-1} + y_{i-1} + y'_{i-1}h + y''(\xi)\frac{h^2}{2}\right] = \\ = y_{i-1}h + y'_{i-1}\frac{h^2}{2} + y''(\xi)\frac{h^3}{4}, \quad \xi \in (x_{i-1}, x_i). \tag{3.55}$$

В соответствии с (3.53) вычтем (3.55) из (3.54), получим погрешность формулы трапеций на одном шаге h:

$$R_i = y''(\xi)h^3\left(\frac{1}{6} - \frac{1}{4}\right) = -\frac{h^3}{12}y''(\xi), \tag{3.56}$$

где $\xi \in (x_{i-1}, x_i)$.

Из рис. 3.11 видно, что если $f''(x) < 0$, то $R_i > 0$, что подтверждается выражением (3.56); если же $f''(x) > 0$, то $R_i < 0$.

На всем отрезке $[a,\, b]$ погрешность (3.56) необходимо увеличить в n раз:

$$R_{\text{тр}} = nR_i = -\frac{nh^3}{12}y''(\xi) = -\frac{(nh)h^2}{12}y''(\xi) = \\ = -\frac{b-a}{12}h^2y''(\xi), \quad \xi \in (a, b). \tag{3.57}$$

Таким образом, метод трапеций — *метод второго порядка точности относительно шага h (главный член погрешности пропорционален шагу в квадрате).*

Поскольку положение точки ξ на интервале (a, b) неизвестно, то, задавая точность ε численного интегрирования, можно записать следующие неравенства, используемые для определения шага h численного интегрирования:

$$|R_{\text{тр}}| \leqslant \max_{x\in[a,b]} |f''(x)| \ \frac{b-a}{12}\ h^2 \leqslant \varepsilon,$$

откуда

$$h \leqslant \sqrt{\frac{12 \cdot \varepsilon}{(b-a)M_2}}, \quad M_2 = \max_{x \in [a,b]} \left| f''(x) \right|. \tag{3.58}$$

Итак, определяющими формулами метода трапеций являются выражения (3.52), (3.57), (3.58).

Численное интегрирование по методу трапеций в случае заданной точности ε осуществляется следующим образом:

1) по формуле (3.58) определяется шаг численного интегрирования h;

2) с помощью этого шага составляется сеточная функция $y_i = f(x_i)$, $i = \overline{0, n}$, для подынтегральной функции $f(x)$ интеграла (3.44);

3) вычисляется приближенное значение интеграла по формуле (3.52), шаг h в которой гарантирует заданную точность ε. Если точность ε не задана, то выбирая шаг h численного интегрирования, можно по формуле (3.57) оценить погрешность $R_{\text{тр}}$ формулы трапеций.

3.5.3. Формула Симпсона численного интегрирования. Разобьем отрезок $[a, b]$ на m пар отрезков $h = x_i - x_{i-1} = \text{const}$, $i = \overline{1, n}$, и через каждые три узла проведем интерполяционный многочлен $L_2(x)$ (рис. 3.12).

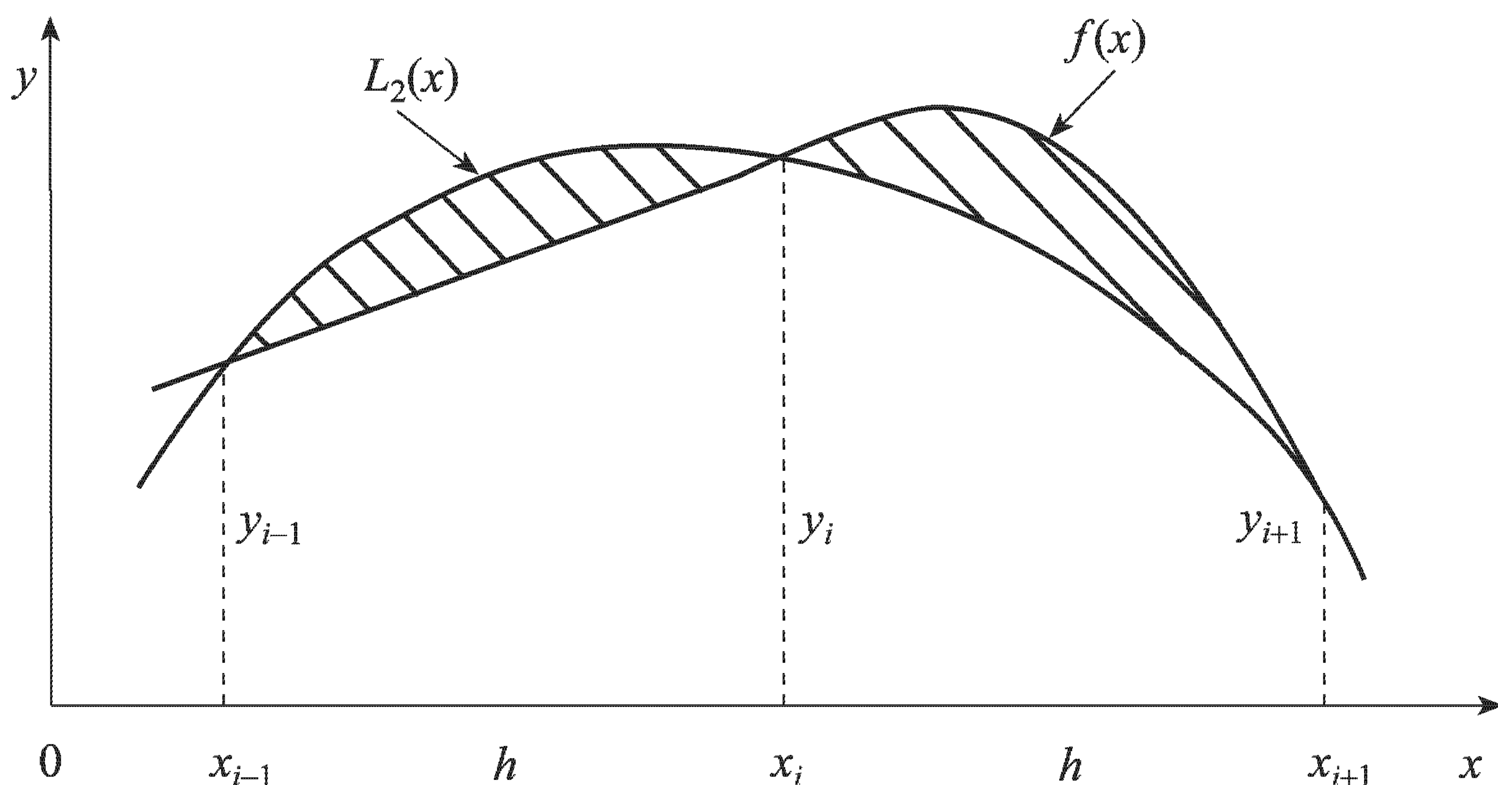


Рис. 3.12. К методу Симпсона численного интегрирования

Тогда

$$\int\limits_{x_{i-1}}^{x_{i+1}} f(x)dx = \int\limits_{x_{i-1}}^{x_{i+1}} L_2(x)dx + R_i,$$

где

$$L_2(x) = y_{i-1}\frac{(x-x_i)(x-x_{i+1})}{(x_{i-1}-x_i)(x_{i-1}-x_{i+1})} + \\ + y_i\frac{(x-x_{i-1})(x-x_{i+1})}{(x_i-x_{i-1})(x_i-x_{i+1})} + y_{i+1}\frac{(x-x_{i-1})(x-x_i)}{(x_{i+1}-x_{i-1})(x_{i+1}-x_i)}.$$

Сделаем замену

$$\frac{x-x_{i-1}}{h} = t, \quad dx = hdt,$$

тогда

$$\frac{x-x_i}{h} = \frac{(x-x_{i-1})-(x_i-x_{i-1})}{h} = t-1;$$

$$\frac{x-x_{i+1}}{h} = \frac{(x-x_{i-1})-(x_{i+1}-x_{i-1})}{h} = t-2.$$

Слагаемые в $L_2(x)$ примут вид

$$y_{i-1}\frac{(x-x_i)(x-x_{i+1})}{h\cdot 2h} = (t-1)(t-2)\frac{y_{i-1}}{2},$$

$$y_i\frac{(x-x_{i-1})(x-x_{i+1})}{-h\cdot h} = -t(t-2)y_i,$$

$$y_{i+1}\frac{(x-x_{i-1})(x-x_i)}{2h\cdot h} = \frac{t}{2}(t-1)y_{i+1}.$$

При $x = x_{i-1}$: $t = 0$; при $x = x_{i+1}$: $t = 2$.

Тогда

$$\int\limits_{x_{i-1}}^{x_{i+1}} L_2(x)dx = h\int\limits_0^2 \left[\frac{y_{i-1}}{2}(t-1)(t-2) - y_i t(t-2) + \right. \\ \left. + \frac{y_{i+1}t}{2}(t-1)\right]dt = \frac{h}{3}(y_{i-1}+4y_i+y_{i+1}),$$

откуда

$$\int\limits_{x_{i-1}}^{x_{i+1}} f(x)dx \approx \int\limits_{x_{i-1}}^{x_{i+1}} L_2(x)dx = \frac{h}{3}(y_{i-1}+4y_i+y_{i+1}). \qquad (3.59)$$

Выражение (3.59) называют формулой Симпсона численного интегрирования на паре шагов от x_{i-1} до x_{i+1}.

На всем отрезке $[a,\ b]$ выражение (3.59) необходимо сложить m раз, поскольку имеется m пар отрезков длиной h, получим *формулу Симпсона численного интегрирования определенного интеграла* (3.44):

$$\int\limits_a^b f(x)dx \approx \frac{h}{3}\left(y_0 + y_n + 4\sum_{i=1}^{m} y_{2i-1} + 2\sum_{i=1}^{m-1} y_{2i}\right). \tag{3.60}$$

Погрешность формулы Симпсона на одной паре шагов записывается следующим образом:

$$R_i = \int\limits_{x_{i-1}}^{x_{i+1}} f(x)dx - \frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}). \tag{3.61}$$

Будем вычислять погрешность в предположении, что $f(x) \in \in C^4$, а $F(x) \in C^5$, где $F'(x) = f(x)$.

Перепишем погрешность (3.61) в виде

$$R_i = \int\limits_{x_i-h}^{x_i+h} f(x)dx - \frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}). \tag{3.62}$$

Тогда на основе формулы Ньютона-Лейбница можно записать

$$\int\limits_{x_i-h}^{x_i+h} f(x)dx = F(x_i + h) - F(x_i - h). \tag{3.63}$$

Разложим значения преобразованных $F(x_i + h)$, $F(x_i - h)$ в ряды Тейлора в окрестности точки x_i до пятой производной включительно:

$$F(x_i + h) = F(x_i) + F'(x_i)h + F''(x_i)\frac{h^2}{2} + F'''(x_i)\frac{h^3}{6} + \\ + F^{IV}(x_i)\frac{h^4}{24} + F^{V}(\xi_1)\frac{h^5}{120}, \quad \xi_1 \in (x_i, x_{i+1});$$

$$F(x_i - h) = F(x_i) - F'(x_i)h + F''(x_i)\frac{h^2}{2} - F'''(x_i)\frac{h^3}{6} +$$

$$+ F^{IV}(x_i)\frac{h^4}{24} - F^V(\xi_2)\frac{h^5}{120}, \quad \xi_2 \in (x_{i-1}, x_i).$$

Тогда, в соответствии с (3.63), получим

$$\int\limits_{x_i-h}^{x_i+h} f(x)dx = 2F'(x_i)h + 2F'''(x_i)\frac{h^3}{6} + [F^V(\xi_1) + F^V(\xi_2)]\frac{h^5}{120}.$$

Для непрерывных на отрезке $[x_{i-1}, x_{i+1}]$ функций найдется такое $\xi \in (\xi_1, \xi_2)$, что

$$F^V(\xi_1) + F^V(\xi_2) = 2F^V(\xi),$$

т. е.

$$\int\limits_{x_i-h}^{x_i+h} f(x)dx = 2hy_i + \frac{2h^3}{6}y_i'' + \frac{h^5}{60}y^{IV}(\xi), \quad \xi \in (x_{i-1}, x_{i+1}). \tag{3.64}$$

Аналогично, разложим y_{i-1} и y_{i+1} в окрестности точки x_i в ряд Тейлора до производных 4-го порядка включительно, получим

$$y_{i-1} = y(x_i - h) = y(x_i) - y'(x_i)h + y''(x_i)\frac{h^2}{2} - y'''(x_i)\frac{h^3}{6} + \\ + y^{IV}(\xi_3)\frac{h^4}{24}, \quad \xi_3 \in (x_{i-1}, x_i),$$

$$y_{i+1} = y(x_i + h) = y(x_i) + y'(x_i)h + y''(x_i)\frac{h^2}{2} + y'''(x_i)\frac{h^3}{6} + \\ + y^{IV}(\xi_4)\frac{h^4}{24}, \quad \xi_4 \in (x_i, x_{i+1});$$

$$\frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}) = \frac{h}{3}\left(6y_i + y_i''h^2 + \frac{h^4}{24}\left[y^{IV}(\xi_3) + y^{IV}(\xi_4)\right]\right),$$

или

$$\frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}) = \\ = 2y_ih + \frac{h^3}{3}y_i'' + \frac{h^5}{36}y^{IV}(\xi), \quad \xi \in (x_{i-1}, x_{i+1}). \tag{3.65}$$

В соответствии с (3.62), (3.64) и (3.65) получим погрешность формулы Симпсона на двойном шаге, которая пропорциональна 4-ой производной подынтегральной функции и пятой степени шага h:

$$R_i = \int\limits_{x_{i-1}}^{x_{i+1}} f(x)dx - \frac{h}{3}(y_{i-1} + 4y_i + y_{i+1}) =$$
$$= -\frac{h^5}{90} y^{IV}(\xi), \quad \xi \in (x_{i-1}, x_{i+1}).$$

Для всего отрезка $[a, b]$ эту погрешность необходимо умножить на m пар отрезков:

$$R_c = mR_i = -\frac{mh^5}{90} f^{IV}(\xi) = -\frac{2mh^5}{180} f^{IV}(\xi) =$$
$$= -\frac{nh \cdot h^4}{180} f^{IV}(\xi) = -\frac{(b-a)h^4}{180} f^{IV}(\xi), \quad \xi \in (a, b), \qquad (3.66)$$

т. е. в формуле Симпсона на всем отрезке $[a, b]$ *погрешность пропорциональна четвертой степени шага, и, следовательно, метод Симпсона является методом четвертого порядка точности* (*т. е. главный член погрешности пропорционален четвертой степени шага* h).

Поскольку положение точки ξ на отрезке $[a, b]$ неизвестно, то в соответствии с (3.66) можно записать верхнюю оценку погрешности и при заданной точности ε получить

$$|R_c| \leqslant \frac{(b-a)h^4}{180} M_4 \leqslant \varepsilon, \quad M_4 = \max_{x \in [a,b]} \left| f^{IV}(x) \right|,$$

откуда

$$h \leqslant \sqrt[4]{\frac{180\varepsilon}{(b-a)M_4}}, \quad M_4 = \max_{x \in [a,b]} \left| f^{IV}(x) \right|. \qquad (3.67)$$

Таким образом, определяющими формулами метода Симпсона являются выражения (3.60), (3.66), (3.67), в соответствии с которыми по заданной точности ε из (3.67) находится шаг h численного интегрирования, с его помощью составляется сеточная функция $y_i = f(x_i)$, $i = \overline{0, n}$, $n = 2m$, а затем приближенно вычисляется интеграл по формуле (3.60). Если точность ε неизвестна, то, задаваясь шагом h, можно по формуле (3.66) вычислить погрешность численного интегрирования.

3.5.4. Процедура Рунге оценки погрешности и уточнения формул численного интегрирования. Процедура Рунге позволяет оценить погрешность и повысить на единицу порядок метода путем многократного (в простейшем случае двукратного) просчета с различными шагами.

Пусть используется какой-либо метод численного интегрирования с шагами h и $h/2$. И пусть порядок выбранного метода равен p, тогда

$$I = I_h + \psi h^p + O(h^{p+1}), \tag{3.68}$$

$$I = I_{h/2} + \psi \left(\frac{h}{2}\right)^p + O(h^{p+1}), \tag{3.69}$$

где I — точное значение интеграла; I_h, $I_{h/2}$ — вычисленные значения интеграла с шагом h и $h/2$ соответственно; вторые слагаемые справа — главные члены погрешности метода численного интегрирования порядка p. Для их вычисления вычтем из выражения (3.69) выражение (3.68), получим

$$\left(I_{h/2} - I_h\right) + \psi \left(\frac{h}{2}\right)^p [1 - 2^p] + O\left(h^{p+1}\right) = 0,$$

$$\psi \left(\frac{h}{2}\right)^p = \frac{I_{h/2} - I_h}{2^p - 1} + O\left(h^{p+1}\right). \tag{3.70}$$

Выражение (3.70) позволяет провести апостериорную оценку погрешности вычисленного значения определенного интеграла.

Подставим (3.70) в (3.69), получим формулу численного интегрирования уже порядка $p+1$:

$$I = I_{h/2} + \frac{I_{h/2} - I_h}{2^p - 1} + O\left(h^{p+1}\right). \tag{3.71}$$

Таким образом, формула (3.71) — простейшая процедура Рунге уточнения на один порядок формулы численного интегрирования.

Замечание. Если для подынтегральной функции $y = f(x)$ построена сеточная функция $y_i = f(x_i)$ с переменным шагом h_i, то погрешность численного интегрирования определяется как интеграл от погрешности интерполяционного многочлена.

Пример 3.8. Методом трапеций с точностью $\varepsilon = 10^{-2}$ и Симпсона с точностью $\varepsilon_1 = 10^{-4}$ вычислить определенный интеграл (вычисляемый точно):

$$\int\limits_0^1 \frac{dx}{1+x} = \ln|1+x|\Big|_0^1 = \ln 2 = 0{,}69315.$$

Решение. 1) *Метод трапеций.* Исходя из заданной точности $\varepsilon = 10^{-2}$ вычислим шаг численного интегрирования, для чего используется формула (3.58):

$$h \leqslant \sqrt{\frac{12\varepsilon}{(b-a)\,M_2}}, \quad M_2 = \max_{x\in[0;\,1]} \left|f''(x)\right| = \max_{x\in[0;\,1]} \left|\frac{2}{(1+x)^3}\right| = 2;$$

$$h \leqslant \sqrt{\frac{12\cdot 0{,}01}{(1-0)\cdot 2}} = \sqrt{6}\cdot 0{,}1 = 0{,}2449.$$

Необходимо выбрать такой шаг, который удовлетворяет неравенству $h \leqslant 0{,}2449$, и чтобы на отрезке интегрирования $x \in [0;\,1]$ он укладывался целое число раз. Принимаем шаг $h = 0{,}2$. Он удовлетворяет обоим этим требованиям.

Для подынтегральной функции $f(x) = (1+x)^{-1}$ с независимой переменной x_i, изменяющейся в соответствии с равенством $x_i = x_0 + ih = 0 + i\cdot 0{,}2$, $i = \overline{0,\,5}$, составляем сеточную функцию с точностью до второго знака после запятой:

i	0	1	2	3	4	5
x_i	0	0,2	0,4	0,6	0,8	1,0
y_i	1,0	0,83	0,71	0,63	0,56	0,5

Используется формула трапеций (3.52) численного интегрирования ($n = 5$):

$$\int\limits_0^1 \frac{dx}{1+x} \approx \frac{h}{2}\left(y_0 + y_n + 2\sum_{i=1}^{n-1} y_i\right) =$$

$$= \frac{0{,}2}{2}\left[1{,}0 + 0{,}5 + 2\,(0{,}83 + 0{,}71 + 0{,}63 + 0{,}56)\right] = 0{,}696.$$

Сравнивая это значение с точным, видим, что абсолютная погрешность не превышает заданной точности ε: $|0{,}69315 - 0{,}696| < 0{,}01$.

Таким образом, за приближенное значение определенного интеграла по методу трапеций с точностью$\varepsilon = 0{,}01$ принимается значение

$$\int_0^1 \frac{dx}{1+x} \approx 0{,}696.$$

2) *Метод Симпсона.* Исходя из заданной точности $\varepsilon_1 = 10^{-4}$ вычисляется шаг численного интегрирования для метода Симпсона по формуле (3.67):

$$h \leqslant \sqrt[4]{\frac{180\varepsilon}{(b-a)\,M_4}}, \quad M_4 = \max_{x\in[0;\,1]} \left|f^{IV}(x)\right| = \max_{x\in[0;\,1]} \left|\frac{24}{(1+x)^5}\right| = 24;$$

$$h \leqslant \sqrt[4]{\frac{180\cdot 10^{-4}}{(1-0)\cdot 24}} = 10^{-1}\sqrt[4]{7{,}5} = 0{,}165.$$

Необходимо выбрать такой шаг, чтобы он удовлетворял неравенству $h \leqslant 0{,}165$, и чтобы на отрезке интегрирования $x \in [0;\,1]$ он укладывался *четное число* раз. Принимаем $h = 0{,}1$. С этим шагом для подынтегральной функции $f(x) = (1+x)^{-1}$ формируется сеточная функция с независимой переменной x_i, изменяющейся по закону $x_i = x_0 + ih = 0 + i\cdot 0{,}1$, $i = \overline{0,\,10}$, $n = 10$, $m = 5$, причем значения сеточной функции вычисляются с точностью до четвертого знака после запятой:

i	0	1	2	3	4	5	6	7	8	9	10
x_i	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
y_i	1,0	0,9091	0,8333	0,7692	0,7143	0,6667	0,625	0,5882	0,5556	0,5263	0,5

Используется формула Симпсона (3.60) численного интегрирования ($n = 10$, $m = 5$)

$$\int_0^1 \frac{dx}{1+x} = \frac{h}{3}\left(y_0 + y_n + 4\sum_{i=1}^{m} y_{2i-1} + 2\sum_{i=1}^{m-1} y_{2i}\right) = \frac{0{,}1}{3}\cdot [1{,}0+$$

$$+0{,}5+4\,(y_1+y_3+y_5+y_7+y_9)+2\,(y_2+y_4+y_6+y_8)]=$$

$$=\frac{0{,}1}{3}\,[1{,}5+4\,(0{,}9091+0{,}7692+0{,}6667+0{,}5882+0{,}5263)+$$

$$+2\,(0{,}8333+0{,}7143+0{,}625+0{,}5556)]=$$

$$=\frac{0{,}1}{3}\,(1{,}5+4\cdot 3{,}4595+2\cdot 2{,}7281)=\frac{0{,}1}{3}20{,}7942=0{,}69314.$$

Сравнение этого значения с точным значением интеграла показывает, что абсолютная погрешность не превышает заданной точности ε_1:

$$|0{,}69315-0{,}69314|<0{,}0001.$$

Таким образом, за приближенное значение определенного интеграла по методу Симпсона с точностью $\varepsilon_1=0{,}0001$ принимается значение

$$\int\limits_0^1 \frac{dx}{1+x}\approx 0{,}6931.$$

Замечание. Ясно, что для большинства интегралов от непрерывных функций первообразная не вычисляется (однако она существует) и вычисленное приближенное значение сравнивать не с чем, однако шаг численного интегрирования, вычисленный по заданной точности, гарантирует эту точность вычисления.

УПРАЖНЕНИЯ.

3.11. Методом трапеций с точностью $\varepsilon=10^{-2}$ и Симпсона с точностью $\varepsilon=10^{-4}$ вычислить определенные интегралы:

а) $\displaystyle\int\limits_0^{0,5}\frac{dx}{\cos x}$; б) $\displaystyle\int\limits_0^1 \cos\,(x+1)^2 dx$; в) $\displaystyle\int\limits_{-1}^0 2^{x^2} dx$;

г) $\displaystyle\int\limits_0^1 e^{x^2} dx$; д) $\displaystyle\int\limits_0^{\pi/2} \sin x^2 dx$; е) $\displaystyle\int\limits_0^{\pi/2} \cos x^2 dx$;

ж) $\displaystyle\int\limits_0^{0,5} \frac{dx}{\cos x^2}$; з) $\displaystyle\int\limits_0^{1} \operatorname{tg}^2 x\,dx$; и) $\displaystyle\int\limits_0^{1} e^{\sin x} dx$;

к) $\displaystyle\int\limits_0^{1} e^{\cos x} dx$; л) $\displaystyle\int\limits_{0,1}^{1} \operatorname{ctg}^2 x\,dx$; м) $\displaystyle\int\limits_0^{1} \operatorname{tg} x^2 dx$.

Для одного из методов применить процедуру Рунге уточнения формулы численного интегрирования.

3.12. Сравнить по трудоемкости методы трапеций и Симпсона для одной и той же точности ε (трудоемкость определяется по количеству вычислений подынтегральной функции).

ГЛАВА IV

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Программа

Постановка задач Коши для обыкновенных дифференциальных уравнений (ОДУ) и систем ОДУ. Методы Эйлера, Эйлера–Коши, Рунге–Кутта численного решения задач Коши для ОДУ и систем ОДУ. Выбор шага численного интегрирования. Порядок метода, процедура Рунге повышения порядка метода. Постановка краевых задач для ОДУ. Конечно-разностный метод с использованием метода прогонки. Аппроксимация краевых условий, содержащих производные, со вторым порядком. Метод пристрелки с использованием различных методов решения нелинейных уравнений.

В этой главе рассматриваются некоторые наиболее употребительные методы численного решения задач Коши для обыкновенных дифференциальных уравнений (ОДУ) и систем ОДУ, а также некоторые численные методы решения краевых задач для ОДУ.

§ 4.1. Основные определения и постановка задач Коши для обыкновенных дифференциальных уравнений

Определение 1. Уравнение, связывающее одну независимую переменную x, искомую функцию $y(x)$ и ее производные до n-го порядка, называется *обыкновенным дифференциальным уравнением* (*ОДУ*):

$$F(x, y, y', \ldots, y^{(n)}) = 0. \tag{4.1}$$

Определение 2. *Порядком ОДУ* называется порядок старшей производной искомой функции $y(x)$. Уравнение (4.1) — ОДУ n-го порядка.

Определение 3. ОДУ называется *линейным*, если функция F линейна относительно искомой функции и ее производных.

Определение 4. Если искомая функция или ее производные входят в ОДУ не в 1-й степени или под знаком трансцендентной функции, то уравнение называется *нелинейным*.

Определение 5. *Общим интегралом* ОДУ называется функция, связывающая независимую переменную x, искомую функцию $y(x)$ и n постоянных интегрирования:

$$\Phi(y(x), x, c_1, c_2, \ldots, c_n) = 0,$$

т. е. $y(x)$ входит в функцию Φ неявным образом.

Определение 6. *Общим решением* ОДУ называется функция

$$y(x) = \varphi(x, c_1, c_2, \ldots, c_n)$$

независимой переменной x и n постоянных интегрирования, т. е. $y(x)$ является явной функцией.

Для определения постоянных интегрирования, т. е. выделения из бесчисленного множества интегральных кривых единственной, проходящей через заданную точку $M(x_0, y_0, y_1, \ldots, y_{n-1})$ в пространстве R^{n+1}, необходимо задать n дополнительных условий в виде значений производных искомой функции в точке x_0 до порядка $n-1$ включительно. Эти дополнительные условия называют *начальными условиями*:

$$\begin{cases} y(x_0) = y_0, \\ y'(x_0) = y_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ y^{(n-1)}(x_0) = y_{n-1}. \end{cases} \tag{4.2}$$

Задача нахождения решения уравнения (4.1), удовлетворяющего n начальным условиям (4.2), называется *задачей Коши* для ОДУ (4.1).

Если старшая производная искомой функции, входящая в ОДУ, выражается явно через производные меньших порядков (такие ОДУ называются *каноническими*), то задача Коши будет иметь вид ($y(x)$ — искомая функция)

$$\begin{cases} y^{(n)} = f(x, y, y', \dots , y^{(n-1)}), \\ y(x_0) = y_0, \\ y'(x_0) = y_1, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ y^{(n-1)}(x_0) = y_{n-1}. \end{cases} \tag{4.3}$$

Определение 7. *Система ОДУ называется нормальной*, если каждое уравнение, входящее в систему, слева содержит производную первого порядка от соответствующей искомой функции, а правая часть зависит от независимой переменной x и n искомых функций $y_1(x)$, $y_2(x)$, ..., $y_n(x)$:

$$\begin{cases} y_1'(x) = f_1(x, y_1, y_2, \dots , y_n), \\ y_2'(x) = f_2(x, y_1, y_2, \dots , y_n), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ y_n'(x) = f_n(x, y_1, y_2, \dots , y_n). \end{cases}$$

При формулировке *задачи Коши для этой системы ОДУ* к ней необходимо присовокупить n начальных условий (по количеству неизвестных), задаваемых в одной и той же точке x_0:

$$\begin{cases} y_1(x_0) = y_{10}, \\ y_2(x_0) = y_{20}, \\ \dots\dots\dots\dots \\ y_n(x_0) = y_{no}. \end{cases}$$

Порядком нормальной системы называют число уравнений в этой системе.

Задачу Коши (4.3) для ОДУ n-го порядка всегда можно привести к задаче Коши для нормальной системы ОДУ того же порядка. Для этого сделаем обозначения:

$$
\begin{aligned}
&\underline{y'(x)} = \underline{z_1(x)},\\
&y''(x) = \underline{z_1'(x)} = \underline{z_2(x)},\\
&y'''(x) = z_1''(x) = \underline{z_2'(x)} = \underline{z_3(x)},\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
&y^{(n-1)}(x) = \underline{z_{n-2}'(x)} = \underline{z_{n-1}(x)}.
\end{aligned}
$$

В соответствии с этими обозначениями дифференциальное уравнение в задаче (4.3) преобразуется в уравнение

$$
y^{(n)}(x) = \underline{z_{n-1}'(x)} = \underline{f(x, y, z_1, z_2, \dots, z_{n-1})}.
$$

Собирая здесь подчеркнутые равенства и приписывая к ним начальные условия из (4.3), получим следующую *задачу Коши для нормальной системы*:

$$
\begin{cases}
y'(x) = z_1(x),\\
z_1'(x) = z_2(x),\\
z_2'(x) = z_3(x),\\
\dots\dots\dots\dots\dots\\
z_{n-2}'(x) = z_{n-1}(x),\\
z_{n-1}'(x) = f(x, y, z_1, z_2, \dots, z_{n-1}),\\
y(x_0) = y_0,\\
z_1(x_0) = y_1,\\
\dots\dots\dots\dots\\
z_{n-1}(x_0) = y_{n-1}.
\end{cases}
$$

В этой задаче Коши искомыми являются n функций: $y(x)$, $z_1(x)$, $z_2(x)$, ..., $z_{n-1}(x)$.

В соответствии с этим ниже рассматриваются численные методы решения задач Коши для ОДУ 1-го порядка и задач Коши для нормальных систем ОДУ.

§ 4.2. Метод Эйлера численного решения задач Коши для ОДУ и систем ОДУ

Пусть дана задача Коши для ОДУ 1-го порядка :

$$\begin{cases} y' = f(x, y), & (4.4)\\ y(x_0) = y_0. & (4.5) \end{cases}$$

Заметим, что правая часть ОДУ (4.4) $f(x, y)$ равна производной y' от искомой функции, и, следовательно, $dy = y'dx = f(x, y)dy$. Для конечно-разностного приращения Δy функции на шаге Δx в точке x_i имеет место равенство $\Delta y_i = f(x_i, y_i)\Delta x$.

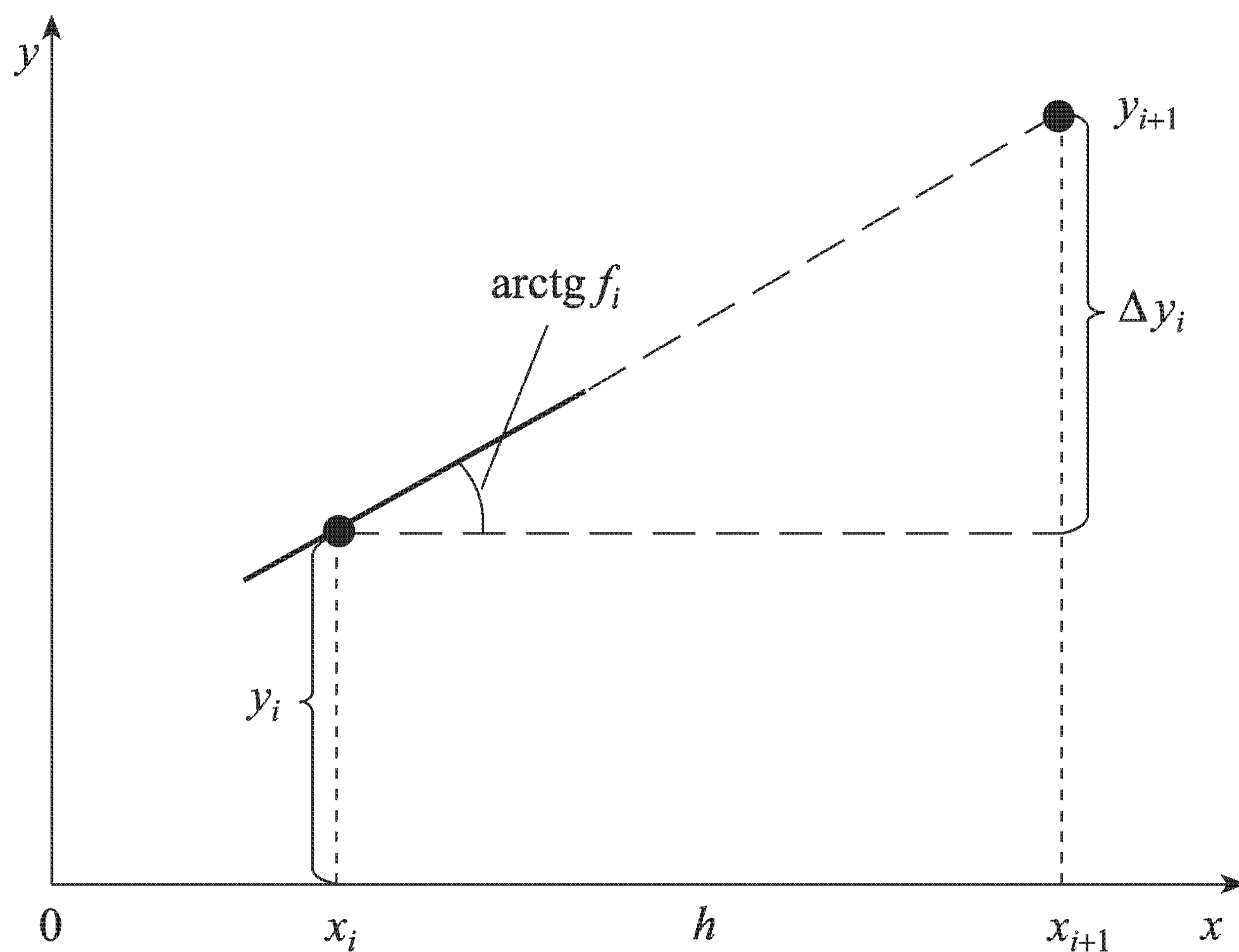


Рис. 4.1. Геометрический смысл метода Эйлера

Интегрируя уравнение (4.4) на отрезке $h = x_{i+1} - x_i$, получим

$$y_{i+1} = y_i + \int\limits_{x_i}^{x_{i+1}} f(x, y(x))dx.$$

К интегралу в правой части этого выражения применим формулу прямоугольников:

$$y_{i+1} = y_i + hf(x_i, y_i) + O(h^2),\ y_0 = y(x_0),\ i = 0, 1, 2, \ldots \quad (4.6)$$

Формула (4.6) — *алгоритм метода Эйлера численного интегрирования задачи Коши* (4.4), (4.5). На каждом шаге численного интегрирования метод Эйлера имеет второй порядок погрешности.

На всем интервале численного интегрирования задачи Коши метод Эйлера, как и метод прямоугольников, имеет погрешность, пропорциональную шагу в первой степени и, следовательно, является *методом первого порядка точности*.

Геометрический смысл метода Эйлера: если известно x_i, y_i, то можно найти правую часть $f(x_i, y_i)$ ОДУ (4.4), т. е. производную y'_i от искомой функции в точке x_i (тангенс угла наклона касательной к интегральной кривой, хотя последняя неизвестна). Продолжая касательную до пересечения с линией $x_{i+1} = x_i + h$, получим y_{i+1} (рис. 4.1, на котором введено обозначение $f_i = f(x_i, y_i)$).

4.2.1. Метод Эйлера для нормальных систем ОДУ. Применительно к нормальным системам ОДУ рассмотрим метод Эйлера на примере задачи Коши для системы второго порядка:

$$\begin{cases} y'_1 = f_1(x, y_1, y_2), & (4.7)\\ y'_2 = f_2(x, y_1, y_2), & (4.8)\\ y_1(x_0) = y_{10}, & (4.9)\\ y_2(x_0) = y_{20}. & (4.10) \end{cases}$$

Выбирая шаг h численного интегрирования $h = x_{i+1} - x_i$, запишем алгоритм (4.6) для каждого уравнения системы ОДУ (4.7), (4.8) ($i = 0, 1, 2, \dots$):

$$\begin{cases} y_{1i+1} = y_{1i} + hf_1(x_i, y_{1i}, y_{2i}) + O(h^2), y_{10} = y_1(x_0); & (4.11)\\ y_{2i+1} = y_{2i} + hf_2(x_i, y_{1i}, y_{2i}) + O(h^2), y_{20} = y_2(x_0). & (4.12) \end{cases}$$

Выражения (4.11), (4.12) описывают алгоритм метода Эйлера численного решения задачи Коши для нормальной системы ОДУ (4.7)–(4.10).

Если задана задача Коши для ОДУ n-го порядка , то она сводится к задаче Коши для нормальных систем. Рассмотрим это на примере задачи Коши для $n = 2$.

$$\begin{cases} y'' = f(x, y, y'), \\ y(x_0) = y_0, \\ y'(x_0) = y_1, \end{cases} \quad \leftrightarrow \quad \begin{cases} y' = z, & (4.13) \\ z' = f(x, y, z), & (4.14) \\ y(x_0) = y_0, & (4.15) \\ z(x_0) = y_1. & (4.16) \end{cases}$$

К задаче (4.13)–(4.16) можно применить алгоритм (4.11), (4.12), если обозначить $f_1(x, y, z) \equiv z$, $f_2(x, y, z) \equiv f(x, y, z)$, причем неизвестными функциями в этой системе являются функции $z(x)$, $y(x)$.

§ 4.3. Метод Эйлера–Коши (Эйлера с пересчетом)

Пусть дана задача Коши (4.4), (4.5) для ОДУ 1-го порядка (4.4). Интегрируя (4.4) на отрезке $h = x_{i+1} - x_i$, $i = 0, 1, 2, \ldots$, получим выражение

$$y_{i+1} = y_i + \int\limits_{x_i}^{x_{i+1}} f(x, y(x))dx.$$

Вычислим интеграл в этом выражении с помощью метода трапеций численного интегрирования с погрешностью на каждом шаге, пропорциональной h^3:

$$y_{i+1} = y_i + \frac{h}{2}\left[f(x_i, y_i) + f(x_{i+1}, y_{i+1})\right] + O\left(h^3\right). \qquad (4.17)$$

Если каким-либо образом вычислено приближение для y_{i+1}, то, приняв его за нулевую итерацию и подставив в правую часть выражения (4.17), получим y_{i+1} на 1-й итерации. На этом основан метод Эйлера–Коши, который выглядит следующим образом. Вычислим вначале y_{i+1} по обычному алгоритму Эйлера и обозначим это значение через $\widetilde{y}_{i+1}$. Затем это значение подставим в правую часть выражения (4.17), получим для $i = 0, 1, 2, \ldots$, $\big(y_0 = y(x_0)$, $x_{i+1} = x_i + h\big)$:

$$\begin{cases} \widetilde{y}_{i+1} = y_i + hf(x_i, y_i), & (4.18) \\ y_{i+1} = y_i + \dfrac{h}{2}\left[f(x_i, y_i) + f(x_{i+1}, \widetilde{y}_{i+1})\right] + O\left(h^3\right). & (4.19) \end{cases}$$

Выражения (4.18), (4.19) описывают алгоритм метода Эйлера–Коши, называемый еще методом Эйлера с пересчетом. Он яв-

ляется частным случаем большого класса методов, который называется методами «предиктор-корректор». Например (4.18) — предиктор, (4.19) — корректор.

На рис. 4.2 представлена геометрическая интерпретация этого метода. Здесь введены обозначения $f_i = f(x_i, y_i)$, $\widetilde{f}_{i+1} = f(x_{i+1}, \widetilde{y}_{i+1})$, $f_{\text{ср}} = 1/2\left(f_i + \widetilde{f}_{i+1}\right)$. Параллельные прямые указаны двусторонними стрелками.

Рис. 4.2. Геометрическая интерпретация метода Эйлера–Коши

Для известных x_i, y_i находят значение производной $f(x_i, y_i) \equiv f_i$ в левом узле шага и $\widetilde{y}_{i+1}$ по (4.18). Зная x_{i+1}, $\widetilde{y}_{i+1}$, можно определить значение производной в правом узле шага: $\widetilde{f}_{i+1} \equiv f(x_{i+1}, \widetilde{y}_{i+1})$, с помощью которого определяется y_{i+1} по формуле (4.19).

Метод Эйлера–Коши имеет 3-й порядок погрешности на каждом шаге, а на всем интервале численного интегрирования погрешность пропорциональна квадрату шага, и поэтому метод Эйлера–Коши является методом второго порядка. К (4.19) можно применить обычную итерационную обработку, а именно: полученное значение y_{i+1} подставить в правую часть вместо $\widetilde{y}_{i+1}$ и получить y_{i+1} на следующей итерации.

4.3.1. Метод Эйлера–Коши для нормальных систем. Рассмотрим метод применительно к задаче Коши для нормальных систем ОДУ 2-го порядка (4.7)–(4.10).

Каждый этап алгоритма (4.18), (4.19) используется сразу для всех неизвестных (в данном случае y_1, y_2), при этом $y_{10} = y_1(x_0)$, $y_{20} = y_2(x_0)$, $x_{i+1} = x_i + h$, $i = 0, 1, 2, \ldots$:

$$\begin{cases} \widetilde{y}_{1i+1} = y_{1i} + hf_1(x_i, y_{1i}, y_{2i}), & (4.20)\\ \widetilde{y}_{2i+1} = y_{2i} + hf_2(x_i, y_{1i}, y_{2i}), & (4.21)\\ y_{1i+1} = y_{1i} + \dfrac{h}{2}[f_1(x_i, y_{1i}, y_{2i}) + f_1(x_{i+1}, \widetilde{y}_{1i+1}, \widetilde{y}_{2i+1})], & (4.22)\\ y_{2i+1} = y_{2i} + \dfrac{h}{2}[f_2(x_i, y_{1i}, y_{2i}) + f_2(x_{i+1}, \widetilde{y}_{1i+1}, \widetilde{y}_{2i+1}). & (4.23) \end{cases}$$

Выражения (4.20)–(4.23) — алгоритм *метода Эйлера–Коши решения задачи Коши* (4.7)–(4.10) *для нормальной системы* (4.7), (4.8) *второго порядка.*

§ 4.4. Метод Рунге–Кутта

Метод Рунге–Кутта имеет погрешность, пропорциональную h^5 на каждом шаге численного интегрирования и поэтому является одним из наиболее употребительных методов численного решения задач Коши для ОДУ и систем ОДУ. Рассмотрим задачу Коши (4.4), (4.5) для ОДУ 1-го порядка (4.4). Проинтегрируем ОДУ (4.4) на шаге $h = x_{i+1} - x_i$, $i = 0, 1, 2, \ldots$:

$$y_{i+1} = y_i + \int\limits_{x_i}^{x_{i+1}} f(x, y(x))dx. \tag{4.24}$$

Для вычисления интеграла в (4.24) используем формулу Симпсона численного интегрирования, имеющую на шаге h точность 5-го порядка. Для формулы Симпсона необходимо три узла (пара шагов), а в (4.24) используются только два узла x_i, x_{i+1}. Возьмем промежуточную точку $x_{i+1/2} = x_i + h/2$ (рис. 4.3), тогда

$$y_{i+1} = y_i + \int\limits_{x_i}^{x_{i+1}} f(x, y(x))dx = y_i + \int\limits_{x_{i+1/2}-h/2}^{x_{i+1/2}+h/2} f(x, y(x))\, dx. \tag{4.25}$$

Теперь для интеграла в выражении (4.25) можно использовать формулу Симпсона с шагом $h/2$:

$$y_{i+1} = y_i + \frac{h/2}{3}\left[f(x_i, y_i) + 4f(x_{i+1/2}, y_{i+1/2}) + \right.$$
$$\left. + f(x_{i+1}, y_{i+1})\right] + O\left(h^5\right). \quad (4.26)$$

Выражение (4.26) является основой для построения различных методов Рунге–Кутта численного решения задач Коши.

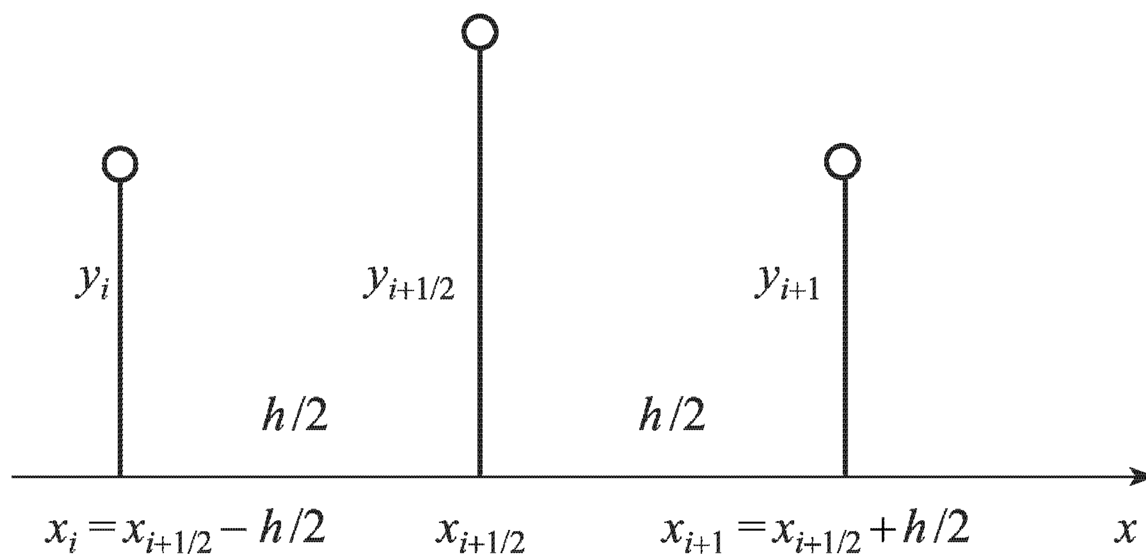


Рис. 4.3. Разбиение шага h в методе Рунге–Кутта

Поскольку выражение (4.26) является неявным (значения y_{i+1} и $y_{i+1/2}$ в правой части неизвестны), то для вычисления второго и третьего слагаемых в правой части полагаем

$$4f(x_{i+1/2}, y_{i+1/2}) + f(x_{i+1}, y_{i+1}) = 2f\left(x_{i+1/2}, y_i + \frac{\Delta y_i^{\mathrm{I}}}{2}\right) +$$
$$+ 2f\left(x_{i+1/2}, y_i + \frac{\Delta y_i^{\mathrm{II}}}{2}\right) + f(x_{i+1}, y_i + \Delta y_i),$$

где приращение Δy_i^{I} вычисляется по x_i, y_i как $\Delta y_i^{\mathrm{I}} = hf(x_i, y_i)$, т. е. как в методе Эйлера; Δy_i^{II} вычисляется по значениям $x_{i+1/2}$, $y_i + \Delta y_i^{\mathrm{I}}/2$; Δy_i — по значениям x_{i+1}, $y_i + \Delta y_i^{\mathrm{II}}$.

В соответствии с такими вычислениями получаем метод Рунге–Кутта 4-го порядка точности, который применяется в следующем виде ($y_0 = y(x_0)$, $x_{i+1/2} = x_i + h/2$, $x_{i+1} = x_i + h$, $i = 0, 1, 2, \ldots$):

$$\begin{cases} y_{i+1} = y_i + \Delta y_i, & (4.27) \\ \Delta y_i = \dfrac{1}{6}[k_i^1 + 2k_i^2 + 2k_i^3 + k_i^4], & (4.28) \end{cases}$$

$$
\begin{cases}
k_i^1 = hf(x_i, y_i), & (4.29)\\
k_i^2 = hf\left(x_{i+1/2}, y_i + \dfrac{k_i^1}{2}\right), & (4.30)\\
k_i^3 = hf\left(x_{i+1/2}, y_i + \dfrac{k_i^2}{2}\right), & (4.31)\\
k_i^4 = hf\left(x_{i+1}, y_i + k_i^3\right). & (4.32)
\end{cases}
$$

Таким образом, метод Рунге–Кутта является четырехэтапным; в нем k_i^1, k_i^2, k_i^3, k_i^4 являются частными приращениями искомой функции соответственно в левом узле x_i шага (4.29), в узле $x_i + h/2$ (4.30) и (4.31) и в правом узле $x_i + h$ шага (4.32). Их линейная комбинация с коэффициентами 1, 2, 2, 1 усредняется в соответствии с (4.28).

Поскольку метод Симпсона на всем интервале численного интегрирования является методом четвертого порядка, то и метод Рунге–Кутта, использующий метод Симпсона, также является *методом четвертого порядка точности.*

Рис. 4.4. Геометрическая интерпретация метода Рунге–Кутта 4-го порядка

На рис. 4.4. представлена геометрическая интерпретация метода Рунге–Кутта. Здесь приняты следующие обозначения:

$$\operatorname{arctg} f_i = \operatorname{arctg} f(x_i, y_i),$$

$$\operatorname{arctg} f^{\mathrm{I}}_{i+1/2} = \operatorname{arctg} f(x_{i+1/2}, y^{\mathrm{I}}_{i+1/2}),$$

$$\operatorname{arctg} f^{\mathrm{II}}_{i+1/2} = \operatorname{arctg} f(x_{i+1/2}, y^{\mathrm{II}}_{i+1/2}),$$

$$\operatorname{arctg} \widetilde{f}_{i+1} = \operatorname{arctg} f(x_{i+1}, \widetilde{y}_{i+1}),$$

$$\operatorname{arctg} f_{\text{ср}} = \frac{1}{6}\left[\operatorname{arctg} f_i + 2\operatorname{arctg} f^{\mathrm{I}}_{i+1/2} + 2\operatorname{arctg} f^{\mathrm{II}}_{i+1/2} + \operatorname{arctg} \widetilde{f}_i\right].$$

На рисунке параллельные прямые указаны двусторонними стрелками.

4.4.1. Метод Рунге–Кутта для нормальных систем ОДУ. Метод Рунге–Кутта для нормальных систем рассмотрим на примере задачи Коши для нормальной системы ОДУ второго порядка (4.7)–(4.10):

$$\begin{cases} y_1' = f_1(x, y_1, y_2), \\ y_2' = f_2(x, y_1, y_2), \\ y_1(x_0) = y_{10}, \\ y_2(x_0) = y_{20}. \end{cases}$$

Будем обозначать частные приращения для искомой функции $y_1(x)$ через k_i^1, k_i^2, k_i^3, k_i^4, а для искомой функции $y_2(x)$ — через l_i^1, l_i^2, l_i^3, l_i^4. Поскольку правые части системы ОДУ $f_1(x, y_1, y_2)$, $f_2(x, y_1, y_2)$ зависят от всех искомых функций (в данном случае от y_1 и y_2), то *приращения для* $y_1(x)$ и $y_2(x)$ *на каждом этапе вычисляются одновременно*. Тогда метод Рунге–Кутта четвертого порядка точности для нормальной системы ОДУ второго порядка примет вид

$$\begin{cases} y_{1i+1} = y_{1i} + \Delta y_{1i}, & (4.33) \\ y_{2i+1} = y_{2i} + \Delta y_{2i}, & (4.34) \\ \Delta y_{1i} = \frac{1}{6}(k_i^1 + 2k_i^2 + 2k_i^3 + k_i^4), & (4.35) \\ \Delta y_{2i} = \frac{1}{6}(l_i^1 + 2l_i^2 + 2l_i^3 + l_i^4), & (4.36) \end{cases}$$

$$
\begin{cases}
k_i^1 = hf_1(x_i, y_{1i}, y_{2i}), & (4.37)\\
l_i^1 = hf_2(x_i, y_{1i}, y_{2i}), & (4.38)\\
k_i^2 = hf_1\left(x_{i+1/2}, y_{1i} + \dfrac{k_i^1}{2}, y_{2i} + \dfrac{l_i^1}{2}\right), & (4.39)\\
l_i^2 = hf_2\left(x_{i+1/2}, y_{1i} + \dfrac{k_i^1}{2}, y_{2i} + \dfrac{l_i^1}{2}\right), & (4.40)\\
k_i^3 = hf_1\left(x_{i+1/2}, y_{1i} + \dfrac{k_i^2}{2}, y_{2i} + \dfrac{l_i^2}{2}\right), & (4.41)\\
l_i^3 = hf_2\left(x_{i+1/2}, y_{1i} + \dfrac{k_i^2}{2}, y_{2i} + \dfrac{l_i^2}{2}\right), & (4.42)\\
k_i^4 = hf_1(x_{i+1}, y_{1i} + k_i^3, y_{2i} + l_i^3), & (4.43)\\
l_i^4 = hf_2(x_{i+1}, y_{1i} + k_i^3, y_{2i} + l_i^3), & (4.44)
\end{cases}
$$

где $x_{i+1/2} = x_i + h/2$, $x_{i+1} = x_i + h$.

Алгоритм (4.33)–(4.44) метода Рунге–Кутта для нормальных систем ОДУ 2-го порядка легко распространяется на нормальные системы 3-го, 4-го и т. д. порядков. Например, для нормальной системы 3-го порядка имеются три неизвестные функции $y_1(x)$, $y_2(x)$, $y_3(x)$, причем для каждой из них вводятся свои частные приращения, а именно: k_i^1, k_i^2, k_i^3, k_i^4 для $y_1(x)$; l_i^1, l_i^2, l_i^3, l_i^4 для $y_2(x)$; m_i^1, m_i^2, m_i^3, m_i^4 для $y_3(x)$. Тогда в методе (4.33)–(4.44) к каждой паре формул добавляется еще одна для неизвестной $y_3(x)$.

§ 4.5. Выбор шага численного интегрирования задач Коши

При численном решении задач Коши для ОДУ и систем ОДУ шаг численного решения можно выбирать апостериорно и априорно. В обоих случаях первоначальное значение шага h задается.

При *апостериорном выборе шага* последний изменяется в процессе счета на основе получаемой информации о поведении решения и на основе заданной точности ε.

Пусть ε — заданная точность численного решения, и пусть h — первоначально выбранный шаг. Тогда алгоритм дальнейшего выбора шага следующий.

1. Выбранным методом на отрезке $x \in [x_0, x_1]$, $x_1 = x_0 + h$ решается задача Коши с шагом h с получением значения $y^h_{x=h}$.

2. Тем же методом с шагом $h/2$ решается задача Коши с получением $y^{h/2}_{x=h/2}$ и $y^{h/2}_{x=h}$.

3. Анализируется неравенство

$$\left| y^h_{x=h} - y^{h/2}_{x=h} \right| \leqslant \varepsilon. \tag{4.45}$$

Если неравенство (4.45) удовлетворяется, то значение шага численного интегрирования на следующем шаге увеличивается вдвое по сравнению с первоначально выбранным шагом, т. е. становится равным $2h$, и алгоритм повторяется начиная с п. 1.

4. Если неравенство (4.45) не выполняется, то счет ведется с шагом $h/4$ начиная с отрезка $x \in [x_0, x_0 + h/4]$ и после получения значения $y^{h/4}_{x=h/2}$ анализируется неравенство

$$\left| y^{h/2}_{x=h/2} - y^{h/4}_{x=h/2} \right| \leqslant \varepsilon.$$

Если оно удовлетворяется, то дальнейший счет ведется с шагом $h/2$ и т. д.

При *априорном выборе шага* расчет ведется с первоначально выбранным шагом h с получением функции $[y(x_i)]_h$, $i = 0, 1, 2, \ldots$, и с шагом $h/2$ с получением функции $[y(x_{2i})]_{h/2}$, $i = 0, 1, 2, \ldots$ Затем анализируется неравенство

$$\max_i \left| [y(x_i)]_h - [y(x_{2i})]_{h/2} \right| \leqslant \varepsilon. \tag{4.46}$$

Если оно выполнено, то решение $[y(x_i)]_{h/2}$, $i = 1, 2, 3, \ldots$, принимается за истинное, в противном случае расчет повторяется с шагом $h/4$ и сравниваются по норме (4.46) функции $[y(x_i)]_{h/2}$ и $[y(x_{2i})]_{h/4}$ и т. д.

§ 4.6. Процедура Рунге оценки погрешности и уточнения численного решения задач Коши

У всех рассмотренных методов численного решения задачи Коши порядок погрешности относительно шага h на всем интервале решения на единицу ниже порядка погрешности на одном шаге h.

Определение. *Порядком* метода назовем показатель p степени h^p в главном члене погрешности метода.

В методе Эйлера главный член погрешности на шаге h пропорционален h^2, а на всем интервале пропорционален шагу h. Поэтому метод Эйлера — метод 1-го порядка. По той же причине метод Эйлера–Коши — метод 2-го порядка, метод Рунге–Кутта — метод 4-го порядка (здесь порядок метода в точности совпадает с порядком соответствующей квадратурной формулы численного интегрирования).

Пусть задача Коши решается методом p-го порядка с шагом h с получением численных значений y_h и главным членом погрешности $\varphi(x)h^p$, пропорциональным h^p. Тогда неизвестное точное решение $y(x)$ можно представить в виде

$$y(x) = y_h + \varphi(x)h^p + O\left(h^{p+1}\right). \tag{4.47}$$

Аналогичное равенство с шагом $h/2$ представим следующим образом:

$$y(x) = y_{h/2} + \varphi(x)(h/2)^p + O\left(h^{p+1}\right). \tag{4.48}$$

Определим $\varphi(x)\left(\dfrac{h}{2}\right)^p$ вычитанием (4.47) из (4.48):

$$\varphi(x)\left(\frac{h}{2}\right)^p = \frac{y_{h/2} - y_h}{2^p - 1} + O\left(h^{p+1}\right). \tag{4.49}$$

Выражение (4.49) дает апостериорную оценку погрешности численного решения. Подставляя (4.49) в (4.48), получим уже метод $(p+1)$-порядка:

$$y(x) = y_{h/2} + \frac{y_{h/2} - y_h}{2^p - 1} + O\left(h^{p+1}\right), \tag{4.50}$$

так как главный член погрешности в алгоритме (4.50) пропорционален степени h^{p+1}.

Процедура (4.50) называется процедурой Рунге уточнения численного решения задачи Коши. Для ее применения задачу необходимо решать дважды с шагами h и $h/2$.

Процесс уточнения с применением формулы (4.50) можно применять и дальше, проводя расчеты с шагами $h/4$, $h/8$ и т. д., пока не выполнится условие

$$\max\left|y_{h/2^{k-1}} - y_{h/2^k}\right| \leqslant \varepsilon, \quad k = 1,\ 2,\ \ldots$$

Пример 4.1. Методами Эйлера, Эйлера–Коши и Рунге–Кутта с шагом $h = 0,1$ численно проинтегрировать следующую задачу Коши для ОДУ 1-го порядка до значения $x = 0,2$ включительно (т. е. два шага):

$$y' = x + y, \quad y(0) = 1, \quad h = 0,1.$$

Задача допускает аналитическое решение и для анализа точности численных методов; выпишем его:

$$y(x) = 2e^x - x - 1;$$

$$y(0) = 1; \quad y(0,1) = 1{,}1103442; \quad y(0,2) = 1{,}242806.$$

Метод Эйлера.

$$\begin{cases} y_{i+1} = y_i + h \cdot (x_i + y_i), \ i = 0, 1, 2, \ldots, \\ y_0 = 1, \ x_i = x_0 + ih = 0 + ih; \end{cases}$$

$$\underline{i = 0; \ x_0 = 0; \ y_0 = 1:}$$

$$y_1 = y_0 + h \cdot (x_0 + y_0) = 1 + 0{,}1\,(0 + 1) = 1{,}1;$$

$$\underline{i = 1; \ x_1 = 0{,}1; \ y_1 = 1{,}1:}$$

$$y_2 = y_1 + h \cdot (x_1 + y_1) = 1{,}1 + 0{,}1\,(0,1 + 1{,}1) = 1{,}22;$$

. . . .

Поскольку метод Эйлера является методом первого порядка точности, то только первая цифра после запятой является верной, а вторая — нет. Это подтверждает сравнение y_1 с $y(0{,}1)$ и y_2 с $y(0{,}2)$.

Метод Эйлера–Коши.

$$\begin{cases} \widetilde{y}_{i+1} = y_i + h\,(x_i + y_i), \\ y_{i+1} = y_i + \dfrac{h}{2}\left[(x_i + y_i) + (x_{i+1} + \widetilde{y}_{i+1})\right], \\ y_0 = 1; \ h = 0{,}1; \ x_i = x_0 + ih, \ i = 0, 1, 2, \ldots; \end{cases}$$

$$\underline{i = 0; \ x_0 = 0; \ x_1 = 0{,}1; \ y_0 = 1:}$$

$$\begin{cases} \widetilde{y}_1 = y_0 + h\cdot(x_0+y_0) = 1 + 0{,}1\,(0+1) = 1{,}1, \\ y_1 = y_0 + \dfrac{h}{2}\left[(x_0+y_0)+(x_1+\widetilde{y}_1)\right] = \\ \quad = 1 + \dfrac{0{,}1}{2}\left[(0+1)+(0{,}1+1{,}1)\right] = 1{,}11; \end{cases}$$

$$\underline{i=1;\ x_1=0{,}1;\ \ x_2=0{,}2;\ \ y_1=1{,}11:}$$

$$\begin{cases} \widetilde{y}_2 = y_1 + h\,(x_1+y_1) = 1{,}11 + 0{,}1\,(0{,}1+1{,}11) = 1{,}231; \\ y_2 = y_1 + \dfrac{h}{2}\left[(x_1+y_1)+(x_2+\widetilde{y}_2)\right] = \\ \quad = 1{,}11 + \dfrac{0{,}1}{2}\left[(0{,}1+1{,}11)+(0{,}2+1{,}231)\right] = 1{,}24205; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

Поскольку метод Эйлера–Коши является методом второго порядка точности, то первые две цифры после запятой считаются верными (сравнить y_1 с $y\,(0{,}1)$ и y_2 с $y\,(0{,}2)$).

Метод Рунге–Кутта.

$$\begin{cases} y_{i+1} = y_i + \Delta y_i, \\ \Delta y_i = \dfrac{1}{6}\left(k_i^1 + 2k_i^2 + 2k_i^3 + k_i^4\right), \end{cases}$$

$$k_i^1 = h\cdot f\,(x_i, y_i) = h\,(x_i+y_i)\,,$$

$$k_i^2 = h\cdot f\left(x_{i+\frac{1}{2}}, y_i + \frac{k_i^1}{2}\right) = h\cdot\left[\left(x_i+\frac{h}{2}\right)+\left(y_i+\frac{k_i^1}{2}\right)\right],$$

$$k_i^3 = h\cdot f\left(x_{i+\frac{1}{2}}, y_i + \frac{k_i^2}{2}\right) = h\cdot\left[\left(x_i+\frac{h}{2}\right)+\left(y_i+\frac{k_i^2}{2}\right)\right],$$

$$k_i^4 = h\cdot f\left(x_{i+1}, y_i + k_i^3\right) = h\left[(x_i+h)+\left(y_i+k_i^3\right)\right];$$

$$\underline{i=0;\ x_0=0;\ x_{0+\frac{1}{2}}=0{,}05;\ x_1=0{,}1;\ y_0=1:}$$

$$\begin{cases} y_1 = y_0 + \Delta y_0 = 1 + 0{,}11035 = 1{,}11035, \\ \Delta y_0 = \dfrac{1}{6}\left(k_0^1 + 2k_0^2 + 2k_0^3 + k_0^4\right) = \dfrac{1}{6}\cdot 0{,}6621 = 0{,}11035, \\ k_0^1 = h\cdot(x_0+y_0) = 0{,}1\,(0+1) = 0{,}1; \end{cases}$$

$$\begin{cases} k_0^2 = h \cdot \left[\left(x_0 + \dfrac{h}{2}\right) + \left(y_0 + \dfrac{k_0^1}{2}\right)\right] = \\ \quad = 0{,}1\left[0{,}05 + \left(1 + \dfrac{0{,}1}{2}\right)\right] = 0{,}11, \\ k_0^3 = h \cdot \left[\left(x_0 + \dfrac{h}{2}\right) + \left(y_0 + \dfrac{k_0^2}{2}\right)\right] = \\ \quad = 0{,}1\left[0{,}05 + \left(1 + \dfrac{0{,}11}{2}\right)\right] = 0{,}1105, \\ k_0^4 = h \cdot \left[(x_0 + h) + \left(y_0 + k_0^3\right)\right] = \\ \quad = 0{,}1\,[0{,}1 + (1 + 0{,}1105)] = 0{,}1211; \end{cases}$$

$$\underline{i = 1;\ x_1 = 0{,}1;\ x_{1+\frac{1}{2}} = 0{,}15;\ x_2 = 0{,}2;\ y_1 = 1{,}11035:}$$

$$\begin{cases} y_2 = y_1 + \Delta y_1 = 1{,}11035 + 0{,}13247 = 1{,}24282, \\ \Delta y_1 = \dfrac{1}{6}\left(k_1^1 + 2k_1^2 + 2k_1^3 + k_1^4\right) = \dfrac{1}{6} \cdot 0{,}7948 = 0{,}13247, \\ k_1^1 = h \cdot (x_1 + y_1) = 0{,}1\,(0{,}1 + 1{,}11035) \approx 0{,}12104; \\ k_1^2 = h \cdot \left[\left(x_1 + \dfrac{h}{2}\right) + \left(y_1 + \dfrac{k_1^1}{2}\right)\right] = \\ \quad = 0{,}1\left[\left(0{,}1 + \dfrac{0{,}1}{2}\right) + \left(1{,}11035 + \dfrac{0{,}12104}{2}\right)\right] = 0{,}132087, \\ k_1^3 = h \cdot \left[\left(x_1 + \dfrac{h}{2}\right) + \left(y_1 + \dfrac{k_1^2}{2}\right)\right] = \\ \quad = 0{,}1\left[\left(0{,}1 + \dfrac{0{,}1}{2}\right) + \left(1{,}11035 + \dfrac{0{,}132087}{2}\right)\right] = 0{,}13264, \\ k_1^4 = h \cdot \left[(x_1 + h) + \left(y_1 + k_1^3\right)\right] = \\ \quad = 0{,}1\,[0{,}2 + (1{,}11035 + 0{,}13264)] = 0{,}1443. \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \end{cases}$$

Метод Рунге–Кутта является методом четвертого порядка точности, поэтому цифры в первых четырех разрядах после запятой верны (сравнить y_1 с $y(0{,}1)$ и y_2 с $y(0{,}2)$).

Пример 4.2. Методами Эйлера, Эйлера–Коши и Рунге–Кутта с шагом $h = 0{,}1$ численно проинтегрировать следующую

задачу Коши для нормальной системы второго порядка до значения $x = 0{,}2$ включительно (т. е. два шага):

$$\begin{cases} y_1' = x + 2y_1 + y_2, \\ y_2' = 2x + y_1 + 2y_2, \\ y_1(0) = 1, \\ y_2(0) = 1. \end{cases}$$

Задача допускает аналитическое решение, которое имеет вид

$$y_1(x) = \frac{7}{6}e^{3x} - \frac{1}{2}e^{x} + \frac{1}{3},$$

$$y_2(x) = \frac{7}{6}e^{3x} + \frac{1}{2}e^{x} - x - \frac{2}{3};$$

$$y_1(0) = 1;\ y_1(0{,}1) = 1{,}355583;\ y_1(0{,}2) = 1{,}848438;$$

$$y_2(0) = 1;\ y_2(0{,}1) = 1{,}36075;\ y_2(0{,}2) = 1{,}86984.$$

Метод Эйлера.

$$\begin{cases} y_{1i+1} = y_{1i} + h\left(x_i + 2y_{1i} + y_{2i}\right), \\ y_{2i+1} = y_{2i} + h\left(2x_i + y_{1i} + 2y_{2i}\right), \\ y_{10} = 1,\ y_{20} = 1; \end{cases}$$

$$\underline{i = 0;\ x_0 = 0; y_{10} = 1;\ y_{20} = 1}:$$

$$y_{11} = y_{10} + h\left(x_0 + 2y_{10} + y_{20}\right) = 1 + 0{,}1\left(0 + 2\cdot 1 + 1\right) = 1{,}3;$$

$$y_{21} = y_{20} + h\left(2x_0 + y_{10} + 2y_{20}\right) = 1 + 0{,}1\left(2\cdot 0 + 1 + 2\cdot 1\right) = 1{,}3;$$

$$\underline{i = 1;\ x_1 = 0{,}1;\ y_{11} = 1{,}3;\ y_{21} = 1{,}3}:$$

$$y_{12} = y_{11} + h\cdot\left(x_1 + 2y_{11} + y_{21}\right) = \\ = 1{,}3 + 0{,}1\left(0{,}1 + 2\cdot 1{,}3 + 1{,}3\right) = 1{,}7;$$

$$y_{22} = y_{21} + h\cdot\left(2\cdot x_1 + y_{11} + 2y_{21}\right) = \\ = 1{,}3 + 0{,}1\left(2\cdot 0{,}1 + 1{,}3 + 2\cdot 1{,}3\right) = 1{,}71;$$

. . . .

Метод Эйлера–Коши.

$$\begin{cases} \widetilde{y}_{1\,i+1} = y_{1i} + h\,(x_i + 2y_{1i} + y_{2i})\,, \\ \widetilde{y}_{2\,i+1} = y_{2i} + h\,(2x_i + y_{1i} + 2y_{2i})\,, \end{cases}$$

$$\begin{cases} y_{1\,i+1} = y_{1i} + \dfrac{h}{2}\,[(x_i + 2y_{1i} + y_{2i}) + (x_{i+1} + 2\widetilde{y}_{1\,i+1} + \widetilde{y}_{2\,i+1})]\,, \\ y_{2\,i+1} = y_{2i} + \dfrac{h}{2}\,[(2x_i + y_{1i} + 2y_{2i}) + (2x_{i+1} + \widetilde{y}_{1\,i+1} + 2\widetilde{y}_{2i+1})]\,, \end{cases}$$

$$y_{10} = 1, \quad y_{20} = 1;$$

$$\underline{i = 0;\ x_0 = 0;\ x_1 = 0{,}1;\ y_{10} = 1;\ y_{20} = 1:}$$

$$\begin{cases} \widetilde{y}_{1\,1} = y_{10} + h\,(x_0 + 2y_{10} + y_{20}) = 1 + 0,1\,(0 + 2\cdot 1 + 1) = 1{,}3; \\ \widetilde{y}_{2\,1} = y_{20} + h\,(2x_0 + y_{10} + 2y_{20}) = 1 + 0,1\,(2\cdot 0 + 1 + 2\cdot 1) = 1{,}3; \end{cases}$$

$$\begin{cases} y_{1\,1} = y_{10} + \dfrac{h}{2}\,[(x_0 + 2y_{10} + y_{20}) + (x_1 + 2\widetilde{y}_{1\,1} + \widetilde{y}_{2\,1})] = \\ \qquad = 1 + \dfrac{0{,}1}{2}\,[(0 + 2\cdot 1 + 1) + (0{,}1 + 2\cdot 1{,}3 + 1{,}3)] = 1{,}35; \\ y_{2\,1} = y_{20} + \dfrac{h}{2}\,[(2x_0 + y_{10} + 2y_{20}) + (2x_1 + \widetilde{y}_{1\,1} + 2\widetilde{y}_{21})] = \\ \qquad = 1 + \dfrac{0{,}1}{2}\,[(2\cdot 0 + 1 + 2\cdot 1) + (2\cdot 0{,}1 + 1{,}3 + 2\cdot 1{,}3)] = 1{,}355; \end{cases}$$

$$\underline{i = 1;\ x_1 = 0{,}1;\ x_2 = 0{,}2;\ y_{11} = 1{,}35;\ y_{21} = 1{,}355:}$$

$$\begin{cases} \widetilde{y}_{1\,2} = y_{11} + h\,(x_1 + 2y_{11} + y_{21}) = \\ \qquad = 1{,}35 + 0{,}1\,(0{,}1 + 2\cdot 1{,}35 + 1{,}355) = 1{,}7655; \\ \widetilde{y}_{2\,2} = y_{21} + h\,(2x_1 + y_{11} + 2y_{21}) = \\ \qquad = 1{,}355 + 0{,}1\,(2\cdot 0{,}1 + 1{,}35 + 2\cdot 1{,}355) = 1{,}781; \end{cases}$$

$$y_{1\,2} = y_{11} + \frac{h}{2}\,[(x_1 + 2y_{11} + y_{21}) + (x_2 + 2\widetilde{y}_{1\,2} + \widetilde{y}_{2\,2})] =$$
$$= 1{,}35 + \frac{0{,}1}{2}[(0{,}1 + 2\cdot 1{,}35 + 1{,}355) +$$
$$+ (0{,}2 + 2\cdot 1{,}7655 + 1{,}781)] = 1{,}8334;$$

$$y_{2\,2} = y_{21} + \frac{h}{2}\left[(2x_1 + y_{11} + 2y_{21}) + (2x_2 + \widetilde{y}_{1\,2} + 2\widetilde{y}_{22})\right] =$$
$$= 1{,}355 + \frac{0{,}1}{2}[(2\cdot 0{,}1 + 1{,}35 + 2\cdot 1{,}355) +$$
$$+ (2\cdot 0{,}2 + 1{,}7655 + 2\cdot 1{,}781)] = 1{,}8544;$$

. . . .

Метод Рунге–Кутта.

$$\begin{cases} y_{1\,i+1} = y_{1i} + \Delta y_{1i}, \\ y_{2i+1} = y_{2i} + \Delta y_{2i}, \end{cases}$$

$$\begin{cases} \Delta y_{1i} = \dfrac{1}{6}\left(k_i^1 + 2k_i^2 + 2k_i^3 + k_i^4\right), \\ \Delta y_{2i} = \dfrac{1}{6}\left(l_i^1 + 2l_i^2 + 2l_i^3 + l_i^4\right), \end{cases}$$

$$\begin{cases} k_i^1 = h\cdot\left[x_i + 2y_{1i} + y_{2i}\right], \\ l_i^1 = h\cdot\left[2x_i + y_{1i} + 2y_{2i}\right], \end{cases}$$

$$\begin{cases} k_i^2 = h\left[\left(x_i + \dfrac{h}{2}\right) + 2\left(y_{1i} + \dfrac{k_i^1}{2}\right) + \left(y_{2i} + \dfrac{l_i^1}{2}\right)\right], \\ l_i^2 = h\left[2\left(x_i + \dfrac{h}{2}\right) + \left(y_{1i} + \dfrac{k_i^1}{2}\right) + 2\left(y_{2i} + \dfrac{l_i^1}{2}\right)\right], \end{cases}$$

$$\begin{cases} k_i^3 = h\left[\left(x_i + \dfrac{h}{2}\right) + 2\left(y_{1i} + \dfrac{k_i^2}{2}\right) + \left(y_{2i} + \dfrac{l_i^2}{2}\right)\right], \\ l_i^3 = h\left[2\left(x_i + \dfrac{h}{2}\right) + \left(y_{1i} + \dfrac{k_i^2}{2}\right) + 2\left(y_{2i} + \dfrac{l_i^2}{2}\right)\right], \end{cases}$$

$$\begin{cases} k_i^4 = h\cdot\left[(x_i + h) + 2\left(y_{1i} + k_i^3\right) + \left(y_{2i} + l_i^3\right)\right], \\ l_i^4 = h\cdot\left[2\,(x_i + h) + \left(y_{1i} + k_i^3\right) + 2\left(y_{2i} + l_i^3\right)\right]; \end{cases}$$

$\underline{i = 0;\ x_0 = 0;\ x_{0+1/2} = 0{,}05;\ x_1 = 0{,}1;\ y_{10} = 1;\ y_{20} = 1:}$

$$\begin{cases} y_{1\,1} = y_{10} + \Delta y_{10} = 1 + 0{,}35558 = 1{,}35558; \\ y_{21} = y_{20} + \Delta y_{20} = 1 + 0{,}36073 = 1{,}36073; \end{cases}$$

$$\begin{cases} \Delta y_{10} = \dfrac{1}{6}\left(k_0^1 + 2k_0^2 + 2k_0^3 + k_0^4\right) = \\ \qquad = \dfrac{1}{6}\left(0{,}3 + 2 \cdot 0{,}35 + 2 \cdot 0{,}3578 + 0{,}4179\right) = 0{,}35558; \\ \Delta y_{20} = \dfrac{1}{6}\left(l_0^1 + 2l_0^2 + 2l_0^3 + l_0^4\right) = \\ \qquad = \dfrac{1}{6}\left(0{,}3 + 2 \cdot 0{,}355 + 2 \cdot 0{,}363 + 0{,}4284\right) = 0{,}36073; \end{cases}$$

$$\begin{cases} k_0^1 = h \cdot [x_0 + 2y_{10} + y_{20}] = 0{,}1\,(0 + 2 \cdot 1 + 1) = 0{,}3; \\ l_0^1 = h \cdot [2x_0 + y_{10} + 2y_{20}] = 0{,}1\,(2 \cdot 0 + 1 + 2 \cdot 1) = 0{,}3; \end{cases}$$

$$\begin{cases} k_0^2 = h\left[\left(x_0 + \dfrac{h}{2}\right) + 2\left(y_{10} + \dfrac{k_0^1}{2}\right) + \left(y_{20} + \dfrac{l_0^1}{2}\right)\right] = \\ \qquad = 0{,}1\left[0{,}05 + 2\left(1 + \dfrac{0{,}3}{2}\right) + \left(1 + \dfrac{0{,}3}{2}\right)\right] = 0{,}35; \\ l_0^2 = h\left[2\left(x_0 + \dfrac{h}{2}\right) + \left(y_{10} + \dfrac{k_0^1}{2}\right) + 2\left(y_{20} + \dfrac{l_0^1}{2}\right)\right] = \\ \qquad = 0{,}1\left[2 \cdot 0{,}05 + \left(1 + \dfrac{0{,}3}{2}\right) + 2\left(1 + \dfrac{0{,}3}{2}\right)\right] = 0{,}355; \end{cases}$$

$$\begin{cases} k_0^3 = h\left[\left(x_0 + \dfrac{h}{2}\right) + 2\left(y_{10} + \dfrac{k_0^2}{2}\right) + \left(y_{20} + \dfrac{l_0^2}{2}\right)\right] = \\ \qquad = 0{,}1\left[0{,}05 + 2\left(1 + \dfrac{0{,}35}{2}\right) + \left(1 + \dfrac{0{,}355}{2}\right)\right] = 0{,}3578; \\ l_0^3 = h\left[2\left(x_0 + \dfrac{h}{2}\right) + \left(y_{10} + \dfrac{k_0^2}{2}\right) + 2\left(y_{20} + \dfrac{l_0^2}{2}\right)\right] = \\ \qquad = 0{,}1\left[2 \cdot 0{,}05 + \left(1 + \dfrac{0{,}35}{2}\right) + 2\left(1 + \dfrac{0{,}355}{2}\right)\right] = 0{,}363; \end{cases}$$

$$\begin{cases} k_0^4 = h \cdot \left[(x_0 + h) + 2\left(y_{10} + k_0^3\right) + \left(y_{20} + l_0^3\right)\right] = \\ \qquad = 0{,}1\,[0{,}1 + 2\,(1 + 0{,}3578) + (1 + 0{,}363)] = 0{,}4179; \\ l_0^4 = h \cdot \left[2\,(x_0 + h) + \left(y_{10} + k_0^3\right) + 2\left(y_{20} + l_0^3\right)\right] = \\ \qquad = 0{,}1\,[2 \cdot 0,1 + (1 + 0{,}3578) + 2\,(1 + 0{,}363)] = 0{,}4284; \end{cases}$$

$$\underline{i=1}; \quad \underline{x_1 = 0{,}1};$$

$$\underline{x_{1+1/2} = 0{,}15}; \quad \underline{x_2 = 0{,}2}; \quad \underline{y_{11} = 1{,}35558}; \quad \underline{y_{21} = 1{,}36073}:$$

$$\begin{cases} y_{1\,2} = y_{11} + \Delta y_{11} = 1{,}35558 + 0{,}49283 = 1{,}8484; \\ y_{22} = y_{21} + \Delta y_{21} = 1{,}36073 + 0{,}5090 = 1{,}8698; \end{cases}$$

$$\begin{cases} \Delta y_{11} = \frac{1}{6}\left(k_1^1 + 2k_1^2 + 2k_1^3 + k_1^4\right) = \\ \qquad = \frac{1}{6}(0{,}41719 + 2\cdot 0{,}4853 + 2\cdot 0{,}4958 + 0{,}57756) = \\ \qquad = 0{,}49283; \\ \Delta y_{21} = \frac{1}{6}\left(l_1^1 + 2l_1^2 + 2l_1^3 + l_1^4\right) = \\ \qquad = \frac{1}{6}(0{,}4277 + 2\cdot 0{,}5013 + 2\cdot 0{,}5121 + 0{,}5997) = 0{,}5090; \end{cases}$$

$$\begin{cases} k_1^1 = h\cdot[x_1 + 2y_{11} + y_{21}] = \\ \qquad = 0{,}1\,(0{,}1 + 2\cdot 1{,}35558 + 1{,}36073) = 0{,}41719; \\ l_1^1 = h\cdot[2x_1 + y_{11} + 2y_{21}] = \\ \qquad = 0{,}1\,(0{,}2 + 1{,}35558 + 2\cdot 1{,}36073) = 0{,}4277; \end{cases}$$

$$\begin{cases} k_1^2 = h\left[\left(x_1 + \frac{h}{2}\right) + 2\left(y_{11} + \frac{k_1^1}{2}\right) + \left(y_{21} + \frac{l_1^1}{2}\right)\right] = \\ \qquad = 0{,}1\left[0{,}15 + 2\left(1{,}35558 + \frac{0{,}41719}{2}\right) + \right. \\ \qquad\qquad \left. + \left(1{,}36073 + \frac{0{,}4277}{2}\right)\right] = 0{,}4853; \\ l_1^2 = h\left[2\left(x_1 + \frac{h}{2}\right) + \left(y_{11} + \frac{k_1^1}{2}\right) + 2\left(y_{21} + \frac{l_1^1}{2}\right)\right] = \\ \qquad = 0{,}1\left[2\cdot 0{,}15 + \left(1{,}35558 + \frac{0{,}41719}{2}\right) + \right. \\ \qquad\qquad \left. +2\left(1{,}36073 + \frac{0{,}4277}{2}\right)\right] = 0{,}5013; \end{cases}$$

$$\begin{cases} k_1^3 = h\left[\left(x_1 + \dfrac{h}{2}\right) + 2\left(y_{11} + \dfrac{k_1^2}{2}\right) + \left(y_{21} + \dfrac{l_1^2}{2}\right)\right] = \\ \quad = 0{,}1\left[0{,}15 + 2\left(1{,}35558 + \dfrac{0{,}4853}{2}\right) + \right. \\ \qquad \left. + \left(1{,}36073 + \dfrac{0{,}5013}{2}\right)\right] = 0{,}4958; \\ l_1^3 = h\left[2\left(x_1 + \dfrac{h}{2}\right) + \left(y_{11} + \dfrac{k_1^2}{2}\right) + 2\left(y_{21} + \dfrac{l_1^2}{2}\right)\right] = \\ \quad = 0{,}1\left[2 \cdot 0{,}15 + \left(1{,}35558 + \dfrac{0{,}4853}{2}\right) + \right. \\ \qquad \left. +2\left(1{,}36073 + \dfrac{0{,}5013}{2}\right)\right] = 0{,}5121; \end{cases}$$

$$\begin{cases} k_1^4 = h \cdot \left[(x_1 + h) + 2\left(y_{11} + k_1^3\right) + \left(y_{21} + l_1^3\right)\right] = \\ \quad = 0{,}1\left[0{,}2 + 2\left(1{,}35558 + 0{,}4958\right) + \right. \\ \qquad \left. + \left(1{,}36073 + 0{,}5121\right)\right] = 0{,}57756; \\ l_1^4 = h \cdot \left[2\left(x_1 + h\right) + \left(y_{11} + k_1^3\right) + 2\left(y_{21} + l_1^3\right)\right] = \\ \quad = 0{,}1\left[2 \cdot 0{,}2 + \left(1{,}35558 + 0{,}4958\right) + \right. \\ \qquad \left. +2\left(1{,}36073 + 0{,}5121\right)\right] = 0{,}5997. \end{cases}$$

Таким образом, метод Эйлера для заданной системы ОДУ (решение которой растет по экспоненте) дает погрешность уже в первой цифре после запятой, метод Эйлера–Коши — во второй, и только метод Рунге–Кутта выдерживает теоретическую точность, сохраняя верными четыре цифры после запятой.

УПРАЖНЕНИЯ.

4.1. Методом Эйлера, Эйлера–Коши, Рунге–Кутта с шагом $h = 0{,}1$ решить следующие задачи Коши для ОДУ 1-го порядка, сделав три шага:

а) $\begin{cases} y' = \dfrac{1}{y} + x, \\ y\,(0) = 1; \end{cases}$ б) $\begin{cases} y' = \dfrac{x^2}{x + y}, \\ y\,(1) = 0; \end{cases}$ в) $\begin{cases} y' = y^2 - x, \\ y\,(0) = 0{,}5; \end{cases}$

г) $\begin{cases} y' = \dfrac{\sin y^2}{y^2 + x}, \\ y\,(0) = 1; \end{cases}$ д) $\begin{cases} y' = \dfrac{y}{x+1} + xy^2, \\ y\,(0) = 1; \end{cases}$ е) $\begin{cases} y' = e^{x/y} + x, \\ y\,(0) = 1. \end{cases}$

4.2. Задачи упражнения 4.1 решить с точностью $\varepsilon = 0{,}01$ для методов Эйлера и Эйлера–Коши и с точностью $\varepsilon_1 = 10^{-4}$ для метода Рунге–Кутта, приняв в качестве начального шага $h = 0{,}1$.

4.3. Методом Рунге–Кутта с точностью $\varepsilon = 10^{-4}$ $(h_0 = 0{,}1)$ решить следующие задачи Коши до значения аргумента $x_{\text{кон}} = 1{,}0$, сведя ОДУ 2-го порядка к нормальной системе ОДУ 2-го порядка:

а) $\begin{cases} y'' = \dfrac{1}{y} + xy', \\ y\,(0) = 1, \\ y'\,(0) = 1; \end{cases}$ б) $\begin{cases} y'' = xy'^2 - y^2, \\ y\,(0) = 1, \\ y'\,(0) = 1; \end{cases}$ в) $\begin{cases} y'' = \dfrac{y'^2}{x^2 + y^2}, \\ y\,(0) = 1, \\ y'\,(0) = 1; \end{cases}$

г) $\begin{cases} y'' = e^{-y'} + xy, \\ y\,(0) = 1, \\ y'\,(0) = 0; \end{cases}$ д) $\begin{cases} y'' = \dfrac{y'}{x+1} + xy^2, \\ y\,(0) = 1, \\ y'\,(0) = 1; \end{cases}$ е) $\begin{cases} y'' = e^{x/y} + x, \\ y\,(0) = 1. \end{cases}$

4.4. В заданиях упражнения 4.1 уточнить решения с помощью процедуры Рунге, просчитав задачи с шагами $h = 0{,}1$; $h_1 = 0{,}05$.

4.5. Методами Эйлера, Эйлера–Коши и Рунге–Кутта с шагом $h = 0{,}1$ до $x_{\text{кон}} = 1$ решить следующие задачи Коши для нормальных систем:

а) $\begin{cases} y_1' = \dfrac{x^2 + y_1^2}{y_2^2}, \\ y_2' = x + y_1 + y_2, \\ y_1\,(0) = 1, \\ y_2\,(0) = 1; \end{cases}$ б) $\begin{cases} y_1' = \ln \dfrac{y_1 + y_2}{x + y_1 y_2}, \\ y_2' = 2x + y_1 - 2y_2, \\ y_1\,(0) = 1, \\ y_2\,(0) = 1; \end{cases}$

в) $\begin{cases} y_1' = \dfrac{x + y_1 + y_2}{x + e^{y_1}}, \\ y_2' = x - 2y_1 + 3y_2, \\ y_1(0) = 1, \\ y_2(0) = 1; \end{cases}$ г) $\begin{cases} y_1' = y_1^2 + e^{xy_2}, \\ y_2' = e^{x+y_1+y_2}, \\ y_1(0) = 1, \\ y_2(0) = 1; \end{cases}$

д) $\begin{cases} y_1' = y_1 e^{-x^2} + xy_2, \\ y_2' = 3x - y_1 + 2y_2, \\ y_1(0) = 1, \\ y_2(0) = 1; \end{cases}$ е) $\begin{cases} y_1' = y_1^2 x + y_2^2, \\ y_2' = \dfrac{2x + y_1}{y_2}, \\ y_1(0) = 1, \\ y_2(0) = 1. \end{cases}$

§ 4.7. Численные методы решения краевых задач для ОДУ

4.7.1. Постановка краевых задач для ОДУ. Рассмотрим постановку краевых задач для ОДУ 2-го порядка с граничными условиями различных родов.

Пусть на отрезке $x \in [a, b]$ определена дважды непрерывно дифференцируемая функция $y(x)$, поведение которой описывается линейным неоднородным ОДУ 2-го порядка. Принципиальным отличием краевой задачи от задачи Коши для ОДУ является задание дополнительных (краевых или граничных) условий более чем в одной точке независимой переменной (в задаче Коши дополнительные условия задаются в одной точке, называемой начальной).

Если на границах $x = a$ и $x = b$ заданы значения искомой функции $y(a)$, $y(b)$, то такие условия называются *граничными условиями первого рода*, а задача

$$\begin{cases} y'' + p(x)y' + q(x)y = f(x), & a < x < b; \quad (4.51) \\ y(a) = y_a, & x = a; \quad (4.52) \\ y(b) = y_b, & x = b \quad (4.53) \end{cases}$$

называется *первой краевой задачей* для ОДУ (4.51).

Если на границах заданы значения производных искомой функции, то такие условия называются *граничными условиями 2-го рода*:

$$y'(a) = y_a; \tag{4.54}$$

$$y'(b) = y_b, \tag{4.55}$$

а задача (4.51), (4.54), (4.55) называется *второй краевой задачей* для ОДУ (4.51).

Если на границах заданы линейные комбинации искомой функции и ее первой производной:

$$y'(a) + \alpha y(a) = y_a, \tag{4.56}$$

$$y'(b) + \beta(b) = y_b, \tag{4.57}$$

то такие условия называются *граничными условиями третьего рода*, а задача (4.51), (4.56), (4.57) называется *третьей краевой задачей* для ОДУ (4.51).

Чаще всего на разных границах задаются граничные условия различных родов. Такие задачи называют краевыми *задачами со смешанными краевыми условиями.*

4.7.2. Конечно-разностный метод с использованием метода прогонки решения краевых задач для ОДУ. Рассмотрим первую краевую задачу (4.51)–(4.53) и будем решать ее конечно-разностным методом, заменяя дифференциальные операторы отношением конечных разностей с использованием формул численного дифференцирования (см. § 3.4).

Для этого введем конечно-разностную сетку с шагом h:

$$\omega_h = \left\{x_i = ih,\ i = \overline{0, n}\right\}.$$

Поскольку ОДУ (4.51) описывает поведение функции $y(x)$ внутри расчетной области $x \in (a, b)$, то производные 1-го и 2-го порядков можно аппроксимировать с помощью отношения центральных разностей со 2-м порядком аппроксимации:

$$y'_i = \frac{y_{i+1} - y_{i-1}}{2h} + O\left(h^2\right),\ i = \overline{1, n-1}; \tag{4.58}$$

$$y''_i = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + O\left(h^2\right),\ i = \overline{1, n-1}; \tag{4.59}$$

$$\{p(x_i), q(x_i), f(x_i)\} \equiv \{p_i, q_i, f_i\},\ i = \overline{1, n-1}. \tag{4.60}$$

Подставляя (4.58)–(4.60) в ОДУ (4.51), получим следующую конечно-разностную схему:

$$\frac{y_{i+1}-2y_i+y_{i-1}}{h^2}+p_i\frac{y_{i+1}-y_{i-1}}{2h}+q_iy_i=f_i+O(h^2),\ i=\overline{1,n-1};$$
$$y_0=y_a,\ i=0;$$
$$y_n=y_b,\ i=n,$$

которую можно представить в виде следующей СЛАУ с трехдиагональной матрицей:

$$a_iy_{i-1}+b_iy_i+c_iy_{i+1}=d_i,\quad i=\overline{1,n-1}, \tag{4.61}$$

где

$$a_i=\frac{1}{h^2}-\frac{p_i}{2h};\quad b_i=-\frac{2}{h^2}+q_i;\quad c_i=\frac{1}{h^2}+\frac{p_i}{2h};\quad d_i=f_i.$$

При $i=1$ первое слагаемое в левой части (4.61) известно и равно $a_1y_0=a_1y_a$; при $i=n-1$ последнее слагаемое в левой части также известно и равно $c_{n-1}y_n=c_{n-1}y_b$. Поэтому СЛАУ (4.61) приобретает следующий вид:

$$a_1=0$$
$$\begin{cases} b_1y_1+c_1y_2=d_1^*,\quad d_1^*=d_1-a_1y_a, & i=1,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_iy_{i-1}+b_iy_i+c_iy_{i+1}=d_i, & i=\overline{2,n-2},\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n-1}y_{n-2}+b_{n-1}y_{n-1}=d_{n-1}^*, & i=n-1,\\ d_{n-1}^*=d_{n-1}-c_{n-1}y_b, & \end{cases} \tag{4.62}$$
$$c_{n-1}=0.$$

Здесь коэффициенты a_1 и c_{n-1} *полагаются равными нулю только после вычисления правых частей* d_1^* и d_{n-1}^*.

Теперь СЛАУ (4.62) пригодна для использования метода прогонки (она имеет трехдиагональную матрицу и $a_1=c_{n-1}=0$).

Прогоночные коэффициенты в прямом ходе определяются с помощью выражений

$$A_i = \frac{-c_i}{b_i + a_i A_{i-1}}; \quad B_i = \frac{d_i - a_i B_{i-1}}{b_i + a_i A_{i-1}}, \quad i = \overline{1, n-1}, \qquad (4.63)$$

причем

$$A_1 = -c_1/b_1, \quad B_1 = d_1/b_1,$$

так как $a_1 = 0$,

$$A_{n-1} = 0,$$

так как $c_{n-1} = 0$.

Значения y_i, $i = n-1, n-2, \ldots, 1$, находятся в обратном ходе с помощью равенств

$$y_i = A_i y_{i+1} + B_i:$$

$$\begin{cases} i = n-1: & y_{n-1} = A_{n-1} y_n + B_{n-1} = B_{n-1}, \\ i = n-2: & y_{n-2} = A_{n-2} y_{n-1} + B_{n-2}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ i = 1: & y_1 = A_1 y_2 + B_1. \end{cases} \qquad (4.64)$$

4.7.3. Конечно-разностная схема со вторым порядком аппроксимации краевых условий, содержащих производные. В случае решения задачи для ОДУ (4.51) с граничными условиями 2-го или 3-го родов на границах $x = a$ и $x = b$ значения искомой функции $y(a)$ и $y(b)$ неизвестны и для их нахождения должны быть составлены алгебраические уравнения в граничных узлах. Однако аппроксимацию производных, входящих в краевые условия, с помощью отношения конечных разностей справа (в узле $x = a$) и слева (в узле $x = b$) можно осуществить только с первым порядком, в то время как дифференциальное уравнение аппроксимируется со вторым порядком. Следовательно, в этом случае вся 2-я (или 3-я) краевая задача будет аппроксимирована только с первым порядком.

Для повышения на единицу порядка аппроксимации производных, входящих в краевые условия, предположим, что искомая функция $y(x)$ дважды дифференцируема не только во внутренних точках расчетной области, но и на границах, т. е. $y(x) \in C^2$, $x \in [a, b]$. Тогда для решения этой проблемы можно использовать

аппарат разложения в ряды Тейлора приграничных значений сеточной функции на точном решении в окрестности граничных узлов.

С этой целью разложим *на точном решении* в ряд Тейлора до 3-й производной включительно y_1 в окрестности узла $x = a$ для левой границы и y_{n-1} в окрестности узла $x = b$ для правой границы.

Затем значения производных 2-го порядка для граничных узлов в этих разложениях заменяются значениями второй производной, определенными из ОДУ, после чего из полученных выражений определяются значения первой производной в граничных узлах со вторым порядком, которые затем подставляются вместо производных первого порядка в краевые условия.

Рассмотрим эту процедуру для левой границы $x_0 = a$:

$$y_1 = y(x_0 + h) = y_0 + y_0' h + y_0'' \frac{h^2}{2} + O(h^3). \tag{4.65}$$

Из (4.51) находим y_0'':

$$y_0'' = f_0 - p_0 y_0' - q_0 y_0. \tag{4.66}$$

Подставив (4.66) в (4.65) и разделив на h, получим

$$y_0' = \frac{y_1 - y_0}{h} - \frac{h}{2}(f_0 - p_0 y_0' - q_0 y_0) + O(h^2),$$

откуда

$$y_0'(1 - \frac{p_0 h}{2}) = \frac{y_1 - y_0}{h} - \frac{h}{2}(f_0 - q_0 y_0) + O(h^2),$$

или

$$y_0' = \frac{2}{2 - p_0 h} \frac{y_1 - y_0}{h} - \frac{h}{2 - p_0 h}(f_0 - q_0 y_0) + O(h^2). \tag{4.67}$$

Подставляя (4.67) в (4.56) (или в (4.54)), получим

$$\frac{2}{2 - p_0 h} \frac{y_1 - y_0}{h} - \frac{h}{2 - p_0 h}(f_0 - q_0 y_0) + \alpha y_0 = y_a + O(h^2). \tag{4.68}$$

Из (4.68) видно, что полученное уравнение для узла $x_0 = a$ содержит только два неизвестных y_0 и y_1, а аппроксимация имеет второй порядок. Следовательно (4.68) можно представить в виде

$$b_0 y_0 + c_0 y_1 = d_0, \tag{4.69}$$

где

$$a_0 = 0; \quad b_0 = -\frac{2}{h(2-p_0h)} + \frac{q_0h}{2-p_0h} + \alpha; \quad c_0 = \frac{2}{h(2-p_0h)};$$

$$d_0 = y_a + \frac{hf_0}{2-p_0h}.$$

Аналогично для правой границы ($x_n = b$):

$$\begin{cases} y_{n-1} = y(x_n - h) = y_n - y'_n h + y''_n \dfrac{h^2}{2} + O(h^3); \\ y''_n = f_n - p_n y'_n - q_n y_n; \end{cases}$$

$$y'_n = \frac{2}{2+p_nh}\frac{y_n - y_{n-1}}{h} + \frac{h}{2+p_nh}(f_n - q_ny_n) + O(h^2).$$

Подставляя это выражение в краевое условие (4.57) (или в (4.55)), получим уравнение для правой границы с двумя неизвестными y_{n-1}, y_n и вторым порядком аппоксимации:

$$\frac{2}{2+p_nh}\frac{y_n - y_{n-1}}{h} + \frac{h}{2+p_nh}(f_n - q_ny_n) + \beta\, y_n = y_b + O(h^2),$$

которое можно представить в виде

$$a_ny_{n-1} + b_ny_n = d_n, \tag{4.70}$$

где

$$a_n = \frac{-2}{h(2+p_nh)}; \quad b_n = \frac{2}{h(2+p_nh)} - \frac{q_nh}{2+p_nh} + \beta; \quad c_n = 0;$$

$$d_n = y_b - \frac{hf_n}{2+p_nh}.$$

Таким образом, результирующая СЛАУ с трехдиагональной матрицей теперь будет содержать $n+1$ уравнение, каждое из которых получено со вторым порядком точности, а именно: уравнение (4.69) при $i = 0$, уравнения (4.61) для $i = \overline{1, n-1}$ и уравнение (4.70) для $i = n$. Для ее решения используется метод прогонки, поскольку $a_0 = 0$ и $c_n = 0$.

4.7.4. Метод пристрелки численного решения краевых задач для ОДУ. Метод пристрелки сводит решение краевой задачи для ОДУ к решению итерационной последовательности задач Коши. Этот метод рассмотрим на примере следующей первой краевой задачи для ОДУ 2-го порядка :

$$\begin{cases} y'' = f(x, y, y'), \quad a < x < b; & (4.71) \\ y(a) = y_0, \qquad x = a; & (4.72) \\ y(b) = y_1, \qquad x = b. & (4.73) \end{cases}$$

Вместо краевой задачи (4.71)–(4.73) рассматривается следующая задача Коши:

$$\begin{cases} y'' = f(x, y, y'), \qquad a < x < b; & (4.74) \\ y(a) = y_0, \qquad x = a; & (4.75) \\ y'(a) = \operatorname{tg} \alpha, \qquad \alpha : \; y(b, \alpha) = y_1, & (4.76) \end{cases}$$

в которой интегральная кривая $y(x, \alpha)$ зависит не только от переменной x, но и от параметра α, который называется *углом*

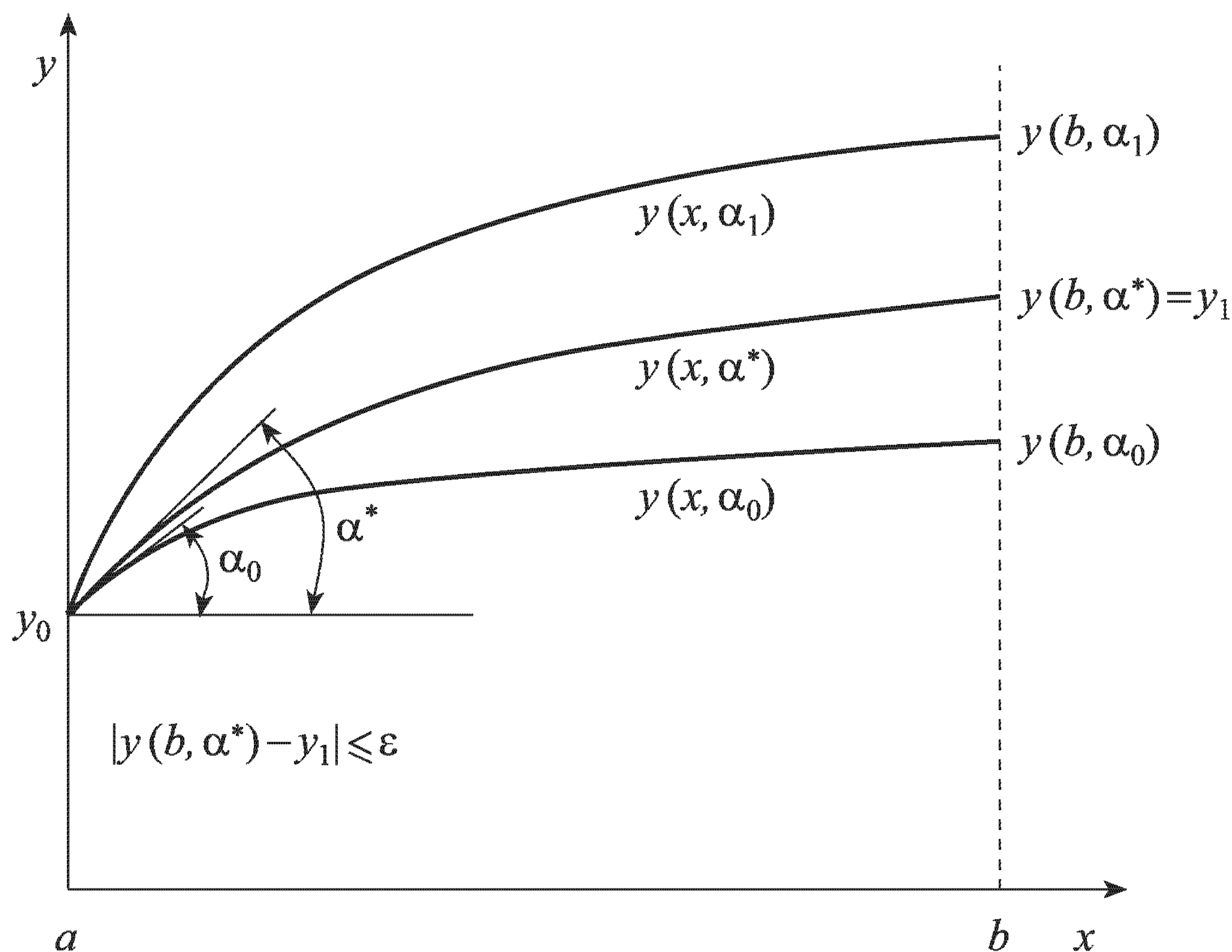


Рис. 4.5. Геометрическая интерпретация метода пристрелки

пристрелки. Он выбирается из условия равенства значения интегральной кривой на правой границе $y(b, \alpha)$ значению y_1 с наперед заданной точностью ε (рис. 4.5):

$$|y(b, \alpha) - y_1| \leqslant \varepsilon. \tag{4.77}$$

Угол пристрелки, удовлетворяющий неравенству (4.77), обозначим через α^*. Интегральная кривая, полученная из решения

задачи Коши (4.74)–(4.76) с углом, близким к этому значению, в соответствии с неравенством (4.77) и будет решением краевой задачи (4.71)–(4.73) с точностью ε.

Таким образом, алгоритм метода пристрелки следующий.

1. Выбирается α_0, например из условия

$$\operatorname{tg}\alpha_0 = \frac{y_1 - y_0}{b - a}.$$

2. С этим значением α_0 одним из методов решается задача Коши (4.74)–(4.76) с получением $y(x, \alpha_0)$ и $y(b, \alpha_0)$; если при этом выполняется условие (4.77), то краевая задача (4.71)–(4.73) решена с точностью ε.

3. В противном случае могут быть следующие два варианта:

а) $y(b, \alpha_0) > y_1$; тогда угол пристрелки каким-либо способом уменьшается и решается задача Коши (4.74)–(4.76) тем же методом до тех пор, пока не выполнится условие $y(b, \alpha_1) < y_1$;

б) $y(b, \alpha_0) < y_1$; тогда угол пристрелки каким-либо способом увеличивается и решается задача Коши до тех пор, пока не выполнится условие $y(b, \alpha_1) > y_1$.

4. Таким образом, угол пристрелки находится внутри интервала $\alpha \in (\alpha_0, \alpha_1)$, после чего истинное значение α^* угла пристрелки определяется методом половинного деления с реализацией следующей цепочки :

а) $\alpha_{k+1} = (\alpha_{k-1} + \alpha_k)/2$;

б) $y(x, \alpha_{k+1})$;

в) $y(b, \alpha_{k+1})$;

г) анализируется неравенство $|y(b, \alpha_{k+1}) - y_1| \leqslant \varepsilon$; если оно выполнено, то $\alpha^* \approx (\alpha_k + \alpha_{k+1})/2$ и $y(x, \alpha^*)$ — истинная интегральная кривая; если неравенство не выполнено, то итерационный процесс повторяется начиная с п. 4 а.

4.7.5. Метод пристрелки с использованием итерационной процедуры Ньютона. Метод половинного деления очень медленно сходится и приходится решать значительное число задач Коши для различных значений углов пристрелки α_k. Для ускорения сходимости итерационного процесса будем полагать, что $y_1 = y(b, \alpha_0 + \Delta\alpha)$, где $\Delta\alpha$ подлежит определению.

Для определения $\Delta\alpha$ разложим $y(b, \alpha_0 + \Delta\alpha)$ в окрестности α_0 в ряд Тейлора до второй производной включительно, получим

$$y(b, \alpha_0 + \Delta\alpha_0) = y(b, \alpha_0) + \frac{\partial y(b, \alpha_0)}{\partial \alpha_0}\Delta\alpha_0 + O(\Delta\alpha^2) = y_1,$$

откуда

$$\Delta\alpha_0 = \frac{y_1 - y(b,\alpha_0)}{\dfrac{\partial y(b,\alpha_0)}{\partial \alpha}}. \tag{4.78}$$

Для определения $\dfrac{\partial y(b,\alpha_0)}{\partial \alpha}$ выбирается малое приращение δ угла пристрелки и со значением угла $\alpha_0 + \delta$ решается задача Коши (4.74)–(4.76) с получением $y(b, \alpha_0 + \delta)$. Тогда для определения производной, входящей в выражение (4.78), используем отношение конечных разностей:

$$\frac{\partial y(b,\alpha_0)}{\partial \alpha} \approx \frac{y(b,\alpha_0+\delta) - y(b,\alpha_0)}{\delta}. \tag{4.79}$$

Подставляя (4.79) в (4.78), получаем

$$\alpha_1 = \alpha_0 + \Delta\alpha_0 = \alpha_0 + \frac{y_1 - y(b,\alpha_0)}{y(b,\alpha_0+\delta) - y(b,\alpha_0)}\delta.$$

Итерационный процесс продолжается до выполнения условия (4.77) с помощью процедуры

$$\alpha_{k+1} = \alpha_k + \frac{y_1 - y(b,\alpha_k)}{y(b,\alpha_k) - y(b,\alpha_{k-1})}(\alpha_k - \alpha_{k-1}),$$

сходящейся значительно быстрее метода половинного деления.

Пример 4.3. Методом конечных разностей с использованием метода прогонки решить следующую краевую задачу для ОДУ 2-го порядка, приняв $h = 0{,}25$ ($a = x_0 = 1$; $b = x_4 = 2$):

$$\begin{cases} y'' = \dfrac{1}{x}y' + x^2, & (а) \\ y'(1) + y(1) = 1, & (б) \\ y(2) = 1. & (в) \end{cases}$$

Решение. Эта задача допускает аналитическое решение, имеющее вид $y(x) = x^4/8 - \dfrac{11}{8}x^2 + 9/2$, значения которого $y(x_i)$ в узлах сетки $x_0 = 1$; $x_1 = 1{,}25$; $x_2 = 1{,}5$; $x_3 = 1{,}75$; $x_4 = 2{,}0$ занесем в таблицу (в эту же таблицу занесем и полученные ниже результаты численного решения y_i, $i = \overline{0,4}$).

i	0	1	2	3	4
x_i	1	1,25	1,5	1,75	2,0
$y(x_i)$	3,25	2,6568	2,039	1,462	1,0
y_i	3,18	2,60	2,00	1,44	1,0

Производные первого и второго порядков, входящие в дифференциальное уравнение, в узлах x_i аппроксимируются со вторым порядком:

$$y_i' = \frac{y_{i+1} - y_{i-1}}{2h} + O\left(h^2\right)\ ; y_i'' = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + O\left(h^2\right).$$

Подставляя эти выражения в дифференциальное уравнение для узлов x_i, $i = \overline{1,\,3}$, получим

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - \frac{1}{x_i}\frac{y_{i+1} - y_{i-1}}{2h} = x_i^2 + O\left(h^2\right), \quad i = \overline{1,\,3}. \quad (г)$$

Поскольку на правой границе ($i = 4,\ x_4 = 2$) задано граничное условие 1-го рода, т. е. значение искомой функции в этом узле известно и равно $y_4 = 1$, то уравнение для этого узла не выписывается (но значение y_4 будет использовано в конечно-разностной аппроксимации (*г*) при $i = 3$). Будем аппроксимировать со вторым порядком левое краевое условие (*б*) в узле $i = 0$, $x_0 = 1$. Разлагая на точном решении y_1 в ряд Тейлора в окрестности граничной точки $x_0 = 1$ до третьей производной включительно, получим

$$y_1 = y\left(x_0 + h\right) = y_0 + y_0'h + y_0''\frac{h^2}{2} + O\left(h^3\right).$$

Подставим сюда вместо y_0'' значение второй производной из дифференциального уравнения (*а*) при $x_0 = 1$: $y_0'' = \dfrac{1}{x_0}y_0' + x_0^2$. Затем разделим все выражение на h и выразим из полученного равенства значение y_0', получим

$$y_0' = \frac{y_1 - y_0}{h\left(1 + \dfrac{h}{2x_0}\right)} - \frac{hx_0^2}{2\left(1 + \dfrac{h}{2x_0}\right)} + O\left(h^2\right).$$

Подставим это выражение в левое краевое условие (*б*) вместо производной первого порядка, получим $(y(1) \equiv y_0)$

$$\frac{y_1 - y_0}{h\left(1 + \dfrac{h}{2x_0}\right)} - \frac{hx_0^2}{2\left(1 + \dfrac{h}{2x_0}\right)} + y_0 = 1 + O\left(h^2\right),$$

откуда при $x_0=1$ и $h = 0{,}25$ получаем алгебраическое уравнение в узле $x_0 = 1$ со вторым порядком:

$$-2{,}556y_0 + 3{,}556y_1 = 1{,}111 + O\ (0{,}0625)\,. \qquad (\partial)$$

Приписывая к этому уравнению алгебраические уравнения, полученные из аппроксимации ($г$) для $i = \overline{1,3}$ $\left(x_1 = 1{,}25;\ x_2 = 1{,}5;\ x_3 = 1{,}75;\ h = 0{,}25\right)$, получим следующую СЛАУ с трехдиагональной матрицей:

$$\begin{cases} -2{,}556y_0 + 3{,}556y_1 = 1{,}111 \\ (a_0 = 0;\ b_0 = -2{,}556;\ c_0 = 3{,}556;\ d_0 = 1{,}111)\,; \\ \\ 1{,}1y_0 - 2y_1 + 0{,}9y_2 = 0{,}0977 \\ (a_1 = 1{,}1;\ b_1 = -2;\ c_1 = 0{,}9;\ d_1 = 0{,}0977)\,; \\ \\ 1{,}0831y_1 - 2y_2 + 0{,}9167y_3 = 0{,}141 \\ (a_2 = 1{,}0831;\ b_2 = -2;\ c_2 = 0{,}9167;\ d_2 = 0{,}141)\,; \\ \\ 1{,}071y_2 - 2y_3 = -0{,}737 \\ (a_3 = 1{,}071;\ b_3 = -2;\ c_3 = 0;\ d_3 = -0{,}737)\,, \end{cases}$$

которая решается методом прогонки:

$$A_0 = -\frac{c_0}{b_0} = \frac{-3{,}556}{-2{,}556} = 1{,}3912;$$

$$B_0 = \frac{d_0}{b_0} = \frac{1{,}111}{-2{,}556} = -0{,}4347;$$

$$A_1 = \frac{-c_1}{b_1 + a_1A_0} = \frac{-0{,}9}{-2 + 1{,}1 \cdot 1{,}3912} = 1{,}9162;$$

$$B_1 = \frac{d_1 - a_1B_0}{b_1 + a_1A_0} = \frac{0{,}0977 - 1{,}1\,(-0{,}4347)}{-2 + 1{,}1 \cdot 1{,}3912} = -1{,}226;$$

$$A_2 = \frac{-c_2}{b_2 + a_2A_1} = \frac{-0{,}9167}{-2 + 1{,}0831 \cdot 1{,}9162} = -12{,}091;$$

$$B_2 = \frac{d_2 - a_2 B_1}{b_2 + a_2 A_1} = \frac{0{,}141 - 1{,}0831\,(-1{,}226)}{-2 + 1{,}0831 \cdot 1{,}9162} = 19{,}3765;$$

$$A_3 = 0;$$

$$B_3 = \frac{d_3 - a_3 B_2}{b_3 + a_3 A_2} = \frac{-0{,}737 - 1{,}071 \cdot 19{,}3765}{-2 + 1{,}071 \cdot (-12{,}091)} = 1{,}4375;$$

$$y_3 = A_3 y_4 + B_3 = B_3 = 1{,}4375;$$

$$y_2 = A_2 y_3 + B_2 = -12{,}091 \cdot 1{,}4375 + 19{,}3765 = 2{,}00;$$

$$y_1 = A_1 y_2 + B_1 = 1{,}9162 \cdot 2{,}00 + (-1{,}226) = 2{,}60;$$

$$y_0 = A_0 y_1 + B_0 = 1{,}3912 \cdot 2{,}60 + (-0{,}4347) = 3{,}18.$$

Эти значения занесем в 3-ю строку таблицы под значениями аналитического решения, сравнение с которым показывает, что действительно конечно-разностная аппроксимация (*г*) и (*д*) имеет второй порядок относительно шага ($h^2 = 0{,}0625$):

$$\Delta\,(y)_i = \{|3{,}25 - 3{,}18| = 0{,}07;\ |2{,}6568 - 2{,}60| = 0{,}0568;\\ |2{,}039 - 2{,}00| = 0{,}039;\ |1{,}462 - 1{,}44| = 0{,}022;\ 0\}$$

Для более точного расчета необходимо уменьшить шаг h.

Пример 4.4. Методом пристрелки с использованием алгоритма Эйлера с шагом $h = 0{,}25$ и методом половинного деления и Ньютона уточнения корней нелинейных уравнений решить следующую первую краевую задачу для ОДУ 2-го порядка (точность $\varepsilon = 0{,}05$):

$$\begin{cases} y'' = \dfrac{1}{x} y' + x^2, \\ y\,(1) = 0, \\ y\,(2) = 1. \end{cases}$$

Эта задача допускает следующее аналитическое решение:

$$y\,(x) = \frac{x^4}{8} - \frac{7}{24} x^2 + \frac{1}{6}.$$

Сведем данную краевую задачу к следующей задаче Коши для того же ОДУ:

$$\begin{cases} y'' = \dfrac{1}{x}y' + x^2, \\ y(1) = 0, \\ y'(1) = \operatorname{tg}\alpha, \ \ \alpha = \alpha^* : y(2, \alpha^*) = 1. \end{cases}$$

Уравнение (4.76) для этой задачи, а именно уравнение $f(\alpha) = y(2, \alpha) - 1 = 0$, как нелинейное уравнение относительно угла пристрелки α можно решить одним из рассмотренных ранее итерационных методов с параллельным решением задачи Коши на каждой итерации. Здесь вначале будет использован метод половинного деления, а затем метод Ньютона.

Метод половинного деления.

1-я итерация. Пусть угол пристрелки α_0 определен из соотношения $\operatorname{tg}\alpha_0 = \dfrac{y(2) - y(1)}{2 - 1} = 1$; $\alpha_0 = 45°$. Тогда задача Коши для ОДУ 2-го порядка и соответствующая задача Коши для нормальной системы ОДУ 2-го порядка будут иметь вид

$$\begin{cases} y'' = \dfrac{1}{x}y' + x^2, \\ y(1) = 0, \\ y'(1) = tg\,\alpha_0 = 1, \end{cases} \quad \leftrightarrow \quad \begin{cases} y' = z, \\ z' = \dfrac{1}{x}z + x^2, \\ y(1) = 0, \\ z(1) = 1. \end{cases}$$

Метод Эйлера с шагом $h = 0,25$ для этой системы имеет вид

$$\begin{cases} y_{i+1} = y_i + h \cdot z_i, \\ z_{i+1} = z_i + h\left(z_i/x_i + x_i^2\right), \ \ i = 0, 1, 2, 3, 4. \end{cases}$$

Решая эту систему, получаем

i	0	1	2	3	4
x_i	1,0	1,25	1,5	1,75	2,0
y_i	0	0,25	0,625	1,173	1,953
z_i	1	1,50	2,19	3,1	4,33

$$y\,(2;\ 45^\circ) = 1{,}953.$$

Итак,

$$f^1\,(45^\circ) = y\,(2; 45^\circ) - 1 = 1{,}953 - 1 = 0{,}953 > 0.$$

Следовательно, необходимо уменьшить угол пристрелки α, причем так, чтобы $f\,(\alpha)$ на следующей итерации было меньше нуля. Примем $\alpha_1 = 0$, тогда задача Коши будет иметь вид:

2-я итерация ($\alpha_1 = 0^\circ$).

$$\begin{cases} y'' = \dfrac{1}{x} y' + x^2, \\ y\,(1) = 0, \\ y'\,(1) = \operatorname{tg}\alpha_1 = 0, \end{cases} \quad \leftrightarrow \quad \begin{cases} y' = z, \\ z' = \dfrac{1}{x} z + x^2, \\ y\,(1) = 0, \\ z\,(1) = 0. \end{cases}$$

Методом Эйлера с шагом $h = 0{,}25$,

$$\begin{cases} y_{i+1} = y_i + h \cdot z_i, \\ z_{i+1} = z_i + h\left(z_i/x_i + x_i^2\right), \quad i = 0, 1, 2, 3, 4, \end{cases}$$

получаем

i	0	1	2	3	4
x_i	1,0	1,25	1,50	1,75	2,0
y_i	0	0	0,0625	0,235	0,577
z_i	0	0,25	0,69	1,368	2,33

$$y\,(2;\ 0^\circ) = 0{,}577;$$

$$f^2\,(0^\circ) = y\,(2; 0^\circ) - 1 = 0{,}577 - 1 = -0{,}423 < 0.$$

Следовательно, угол α^* находится внутри интервала $\alpha_1 = 0^\circ < \alpha^* < \alpha_0 = 45^\circ$. После этого для уточнения угла пристрелки α используем метод половинного деления.

3-я итерация $\left(\alpha_2 = \dfrac{\alpha_0 + \alpha_1}{2} = \dfrac{45^\circ + 0^\circ}{2} = 22{,}5^\circ\right)$; $\operatorname{tg} 22{,}5^\circ = 0{,}414$.

$$\begin{cases} y'' = \dfrac{1}{x} y' + x^2, \\ y\,(1) = 0, \\ y'\,(1) = \operatorname{tg}\alpha_2 = 0{,}414, \end{cases} \quad \leftrightarrow \quad \begin{cases} y' = z, \\ z' = \dfrac{1}{x} z + x^2, \\ y\,(1) = 0, \\ z\,(1) = 0{,}414. \end{cases}$$

Решая эту задачу Коши методом Эйлера с шагом $h = 0{,}25$, получим

i	0	1	2	3	4
x_i	1,0	1,25	1,5	1,75	2,0
y_i	0	0,104	0,296	0,624	1,147
z_i	0,414	0,768	1,312	2,093	3,153

$$y\,(2;\ 22{,}5^\circ) = 1{,}147;$$

$$f^3\,(22{,}5^\circ) = y\,(2;\ 22{,}5^\circ) - 1 = 1{,}147 - 1 = 0{,}147 > 0;$$

$$\left|f^3\,(22{,}5^\circ)\right| = 0{,}147 > \varepsilon = 0{,}05.$$

Сравнение значения функции $f\,(\alpha)$ на 2-й итерации $\left(f^2\,(0^\circ) < 0\right)$ и на третьей итерации $\left(f^3\,(22{,}5^\circ) > 0\right)$ означает, что $0^\circ < \alpha^* < 22{,}5^\circ$.

4-я итерация $\left(\alpha_3 = \dfrac{0^\circ + 22{,}5^\circ}{2} = 11{,}25^\circ, \quad \operatorname{tg} 11{,}25^\circ = 0{,}2\right)$.

Решение задачи Коши

$$\begin{cases} y' = z, \\ z' = \dfrac{1}{x} z + x^2, \\ y\,(1) = 0, \\ z\,(1) = 0{,}2 \end{cases}$$

методом Эйлера с шагом $h = 0{,}25$ сведено в таблицу

i	0	1	2	3	4
x_i	1	1,25	1,5	1,75	2,0
y_i	0	0,05	0,175	0,423	0,853
z_i	0,2	0,5	0,991	1,719	2,73

$$y\,(2;\ 11{,}25^\circ) = 0{,}853;$$

$$f^4\,(11{,}25^\circ) = y\,(2;\ 11{,}25^\circ) - 1 = 0{,}853 - 1 = -0{,}147 < 0;$$

$$\left|f^4\,(11{,}25^\circ)\right| = 0{,}147 > \varepsilon = 0{,}05.$$

Сравнение $f\,(\alpha)$ на третьей итерации $\left(f^3\,(22{,}5^\circ) > 0\right)$ и на четвертой $\left(f^4\,(11{,}25^\circ) < 0\right)$ означает, что $11{,}25^\circ < \alpha^* < 22{,}5^\circ$.

5-я итерация $\left(\alpha_4 = \dfrac{11{,}25^\circ + 22{,}5^\circ}{2} = 16{,}875^\circ;\quad \operatorname{tg} 16{,}875^\circ = = 0{,}3\right)$.

Решение задачи Коши

$$\begin{cases} y' = z, \\ z' = \dfrac{1}{x} z + x^2, \\ y\,(1) = 0, \\ z\,(1) = 0{,}3 \end{cases}$$

методом Эйлера с шагом $h = 0{,}25$ сведено в таблицу

i	0	1	2	3	4
x_i	1,0	1,25	1,5	1,75	2,0
y_i	0	0,075	0,23	0,515	0,99
z_i	0,3	0,625	1,14	1,89	2,93

$$y\,(2;\ 16{,}875^\circ) = 0,99;$$

$$f^5\,(16{,}875^\circ) = y\,(2;\ 16{,}875^\circ) - 1 = 0{,}99 - 1 = -0{,}01;$$

$$\left|f^5\,(16{,}875^\circ)\right| = 0{,}01 < 0{,}05.$$

Таким образом, угол пристрелки принимается равным $\alpha^* = 16,875°$. На этом угле интегральная кривая имеет значения y_i из последней таблицы.

Метод Ньютона.

Как видно из проведенного расчета, метод половинного деления при уточнении нуля α^* функции $f(\alpha)$ очень медленно сходится. Для ускорения итерационного процесса применим метод Ньютона, для чего на первой итерации примем $\alpha_0 = 45°$. Результаты расчетов по методу Эйлера с использованием метода половинного деления приведены в таблице первой итерации, из которой следует, что $y(2;\ 45°) = 1,953$. Для проведения 2-й итерации вычислим α_1 так:

$$\alpha_1 = \alpha_0 + \Delta\alpha_0 = \alpha_0 + \frac{y(2) - y(2;\ 45°)}{y(2; 45° + \delta) - y(2;\ 45°)}\delta,$$

где δ примем равным $-10°$ (угол α нужно уменьшать, поскольку $f^1(45°)=0{,}953>0$, а $f(\alpha^*)=0$).

С углом $\alpha = \alpha_0 + \delta = 45° - 10° = 35°$ решается задача Коши методом Эйлера с шагом $h = 0{,}25$. В результате получаем таблицу

i	0	1	2	3	4
x_i	1,0	1,25	1,50	1,75	2,0
y_i	0	0,175	0,456	0,89	1,538
z_i	0,7	1,125	1,74	2,59	3,73

$$y(2;\ 35°) = 1{,}538.$$

На второй итерации

$$\alpha_1 = 45° + \frac{y(2) - y(2; 45°)}{y(2; 35°) - y(2; 45°)} \cdot (-10°) = 22{,}5°;$$

$$\operatorname{tg} 22{,}5° = 0{,}404.$$

С этим α_1 на второй итерации решается задача Коши, результаты решения которой приведены в таблице на 3-й итерации метода половинного деления, откуда $y(2;\ 22{,}5°) = 1{,}147$, а $\left|f^2(22{,}5°)\right| = |1{,}147 - 1| = 0{,}147 > \varepsilon = 0{,}05$.

На 3-й итерации

$$\alpha_2 = \alpha_1 + \frac{y(2) - y(2;\ 22{,}5^\circ)}{y(2;\ 45^\circ) - y(2; 22{,}5^\circ)}(45^\circ - 22{,}5^\circ) =$$

$$= 22{,}5^\circ + \frac{1 - 1{,}147}{1{,}953 - 1{,}147} \cdot (45^\circ - 22{,}5^\circ) = 17{,}8^\circ;$$

$$\operatorname{tg} 17{,}8^\circ = 0{,}32.$$

Решая задачу Коши с этим углом $\alpha_2 = 17{,}8^\circ$, получаем

i	0	1	2	3	4
x_i	1,0	1,25	1,5	1,75	2,04
y_i	0	0,08	0,243	0,535	1,016
z_i	0,32	0,65	1,17	1,93	2,97

$$y(2;\ 17{,}8^\circ) = 1{,}016;$$

$$f^3(17{,}8^\circ) = |y(2; 17{,}8^\circ) - 1| = |1{,}016 - 1| = 0{,}016 < \varepsilon = 0{,}5.$$

Таким образом, точность выполнена, и итерационный процесс останавливается. Кроме этого, видно, что метод Ньютона сходится значительно быстрее метода половинного деления.

Ответом являются значения y_i из этой таблицы.

УПРАЖНЕНИЯ.

4.6. Методом конечных разностей с использованием метода прогонки решить следующие краевые задачи для ОДУ 2-го порядка (аппроксимацию производных в краевых условиях выполнить с $O\left(h^2\right)$, $h = 0{,}1$):

а) $\begin{cases} y'' = x^2 y' + e^x, \\ y'(0) + y(0) = 0, \\ y'(1) = 1; \end{cases}$ б) $\begin{cases} y'' = \sin x \cdot y' - x^2, \\ y'(0) = 0, \\ y(1) = 1; \end{cases}$

в) $\begin{cases} y'' = xy' - e^x, \\ y'(0) + y(0) = 1, \\ y'(1) = 0; \end{cases}$ г) $\begin{cases} y'' = e^x \cdot y' + xy, \\ y'(0) - y(0) = 1, \\ y'(1) + y(1) = 0; \end{cases}$

д) $\begin{cases} y'' = e^x y' + x, \\ y'(0) = 1, \\ y'(1) - y(1) = 0; \end{cases}$ е) $\begin{cases} y'' = xy + 1, \\ y'(0) + y(0) = 1, \\ y(1) = 0; \end{cases}$

ж) $\begin{cases} y'' = 2^x y' + 2, \\ y'(0) + y(0) = 1, \\ y(1) = 0; \end{cases}$ з) $\begin{cases} y'' = -xy' + 1, \\ y'(0) - 2y(0) = 1, \\ y'(1) = 0. \end{cases}$

4.7. Методом пристрелки с использованием метода половинного деления или Ньютона с точностью $\varepsilon = 0,01$ решить следующие краевые задачи для ОДУ 2-го порядка (применить один из методов: Эйлера, Эйлера–Коши или Рунге–Кутта):

а) $\begin{cases} y'' = \cos x \cdot y' - x^2, \\ y(0) = 1, \\ y(1) = 1{,}5; \end{cases}$ б) $\begin{cases} y'' = -xy' + e^x, \\ y(0) = 0, \\ y(1) = 1; \end{cases}$

в) $\begin{cases} y'' = e^x y' + e^{-x}, \\ y(0) = 1, \\ y'(1) = 1; \end{cases}$ г) $\begin{cases} y'' = \sin x \cdot y' - \cos x, \\ y(0) = 0, \\ y'(1) = 1; \end{cases}$

д) $\begin{cases} y'' = x^2 y - e^x, \\ y(0) = 1, \\ y(1) = 0; \end{cases}$ е) $\begin{cases} y'' = e^{-x} y' + x, \\ y(0) = 0, \\ y'(1) = 2; \end{cases}$

Указание: В случае задания граничного условия 2-го рода на правой границе итерационный процесс осуществлять до выполнения условия

$$\left|\frac{y(1)-y(1-h)}{h}-y'(1)\right|\leqslant\varepsilon,$$

причем в качестве начального угла пристрелки можно принять $\alpha_0=\operatorname{arctg} y'(1)$.

ГЛАВА V

ЧИСЛЕННЫЕ МЕТОДЫ ОПТИМИЗАЦИИ

Программа

Численные методы безусловной минимизации функции одной переменной, прямые методы: перебора, половинного деления, золотого сечения. Методы минимизации, использующие производные. Численные методы безусловной минимизации функций многих переменных: градиентного спуска, наискорейшего спуска, сопряженных направлений.

§ 5.1. Классификация численных методов оптимизации

Задачи оптимизации, за редким исключением, решаются с помощью вычислительных процедур, поэтому названия главы «Методы оптимизации» и «Численные методы оптимизации (ЧМО)» практически равносильны. Среди важнейших методов оптимизации широко распространены следующие методы.

1. Безусловной минимизации функций одной переменной:
— прямые методы (перебора, половинного деления, золотого сечения);
— методы, использующие производные первого и второго порядка (Ньютона, ломаных, касательных).

2. Безусловной минимизации функций многих переменных:
— градиентного спуска;
— наискорейшего спуска;
— сопряженных направлений.

3. Условной оптимизации:
— линейное и квадратичное программирование;
— нелинейное программирование;
— динамическое программирование.

Из перечисленных методов будем рассматривать только численные методы безусловной минимизации, поскольку широкий класс методов математического программирования входит в отдельный курс «Методы оптимизации».

§ 5.2. Численные методы безусловной минимизации функций одной переменной. Прямые методы

Под *минимизацией* функции $f(x)$ на множестве $X \subset R$ будем понимать следующую задачу: найти хотя бы одну *точку минимума* x^* и *минимум* $f^* = f(x^*)$ на множестве X.

Задача нахождения точки максимума x^* функции $f(x)$ и максимального значения функции $f(x^*)$ сводится к задаче минимизации функции $-f(x)$, поэтому ниже будем рассматривать только задачу минимизации.

Число $x^* \in X$ называется *точкой абсолютного* (*глобального*) *минимума* функции $f(x)$, а значение $f(x^*)$ — *глобальным минимумом*, если $f(x^*) \leqslant f(x)$ для всех $x \in X$.

Число $\widetilde{x} \in X$ называется *точкой локального минимума*, если существует такое число $\delta > 0$, что выполняется неравенство $f(\widetilde{x}) \leqslant f(x)$ для всех точек x, удовлетворяющих условию $|x - \widetilde{x}| < \delta$; значение $f(\widetilde{x})$ называется *локальным минимумом*.

Будем рассматривать *унимодальные функции* $f(x)$ — функции, имеющие один минимум в области X.

Функция $f(x)$ называется унимодальной на отрезке $[a, b]$, если она непрерывна на $[a, b]$ и существуют числа α и β, $a \leqslant \alpha \leqslant \beta \leqslant b$, такие, что

1) на отрезке $[a, \alpha]$ $f(x)$ монотонно убывает;

2) на отрезке $[\beta, b]$ $f(x)$ монотонно возрастает;

3) на отрезке $x \in [\alpha, \beta]$ $f(x)$ имеет минимум $f^* = \min\limits_{x \in [\alpha, \beta]} f(x)$.

Существует два критерия унимодальности, используемые на практике:

1) если функция $f(x)$ дифференцируема на отрезке $[a, b]$ и производная $f'(x)$ *не убывает* на этом отрезке, то $f(x)$ *унимодальна.*

2) Если функция $f(x)$ дважды дифференцируема на отрезке $[a, b]$ и $f''(x) \geqslant 0$ на этом отрезке, то $f(x)$ *унимодальна.*

Пример 5.1. Показать, что функция $f(x) = x^4 - 10x^3 + 36x^2 + 5x$ унимодальна на отрезке $[3;\ 5]$.

Решение. Вторая производная функции $f(x)$ равна $f''(x) = 12x^2 - 60x + 72$. Корни полученного квадратного трехчлена: $x_1 = 2$ и $x_2 = 3$. Следовательно $f''(x) \geqslant 0$ при $x \in (-\infty, 2] \cup [3, \infty)$. Пересечением этих промежутков с заданным отрезком

будет отрезок [3; 5] и по второму критерию унимодальности заключаем, что $f(x)$ унимодальна на $x \in [3;\ 5]$.

Прямые методы минимизации основаны на вычислении только значений минимизируемой функции и не используют значений ее производных.

Из прямых методов минимизации рассмотрим методы перебора, половинного деления, золотого сечения.

5.2.1. Метод перебора. Метод перебора является простейшим методом из прямых методов минимизации, но он является и самым надежным методом в смысле гарантированного нахождения минимального значения.

Пусть $f(x)$ унимодальна на $x \in [a, b]$. Требуется найти точку минимума x^* и минимум $f(x^*)$ с точностью $\varepsilon > 0$.

Отрезок $[a, b]$ разбивается точками $x_i = a + i \cdot \dfrac{b-a}{n}$, $i = \overline{0, n}$, на n равных частей, где $n \geqslant \dfrac{b-a}{\varepsilon}$ (в простейшем случае $n = \dfrac{b-a}{\varepsilon}$ или $\dfrac{b-a}{n} = \varepsilon$).

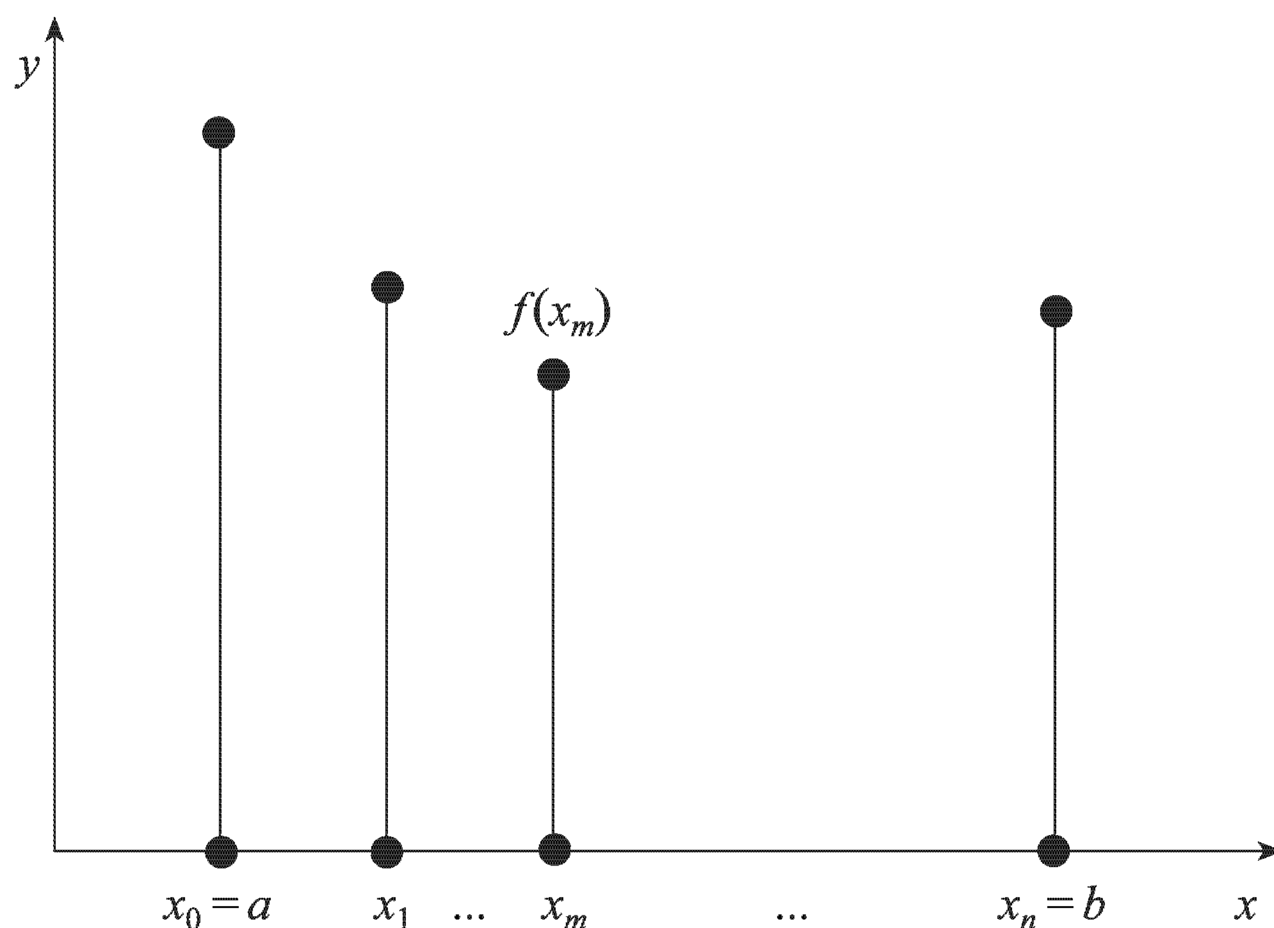


Рис. 5.1. К методу перебора

Вычислив значения $f(x_i)$, $i = \overline{0, n}$, путем сравнения находим (см. рис. 5.1)

$$f(x_m) \approx \min f(x_i); \quad x^* \approx x_m; \quad f^* \approx f(x_m).$$

Ясно, что значения x^* и $f(x^*)$ являются приближенными. Поэтому для вычисления минимума с большей точностью ε_1 $(\varepsilon_1 < \varepsilon)$ рассматривается отрезок $x \in [x_{m-1}, x_{m+1}]$, где находится минимальное значение $f(x^*)$, на нем вычисляются значения $f(x_j)$, $j = \overline{0, k}$, $k = (x_{m+1} - x_{m-1})/\varepsilon_1$, и путем сравнения $f(x_j)$ находим более точное значение минимума функции.

Достоинства метода — простота и надежность, недостаток — большое число вычислений значений функции $f(x)$.

Пример 5.2. Методом перебора с точностью $\varepsilon = 0,05$ найти точку минимума x^* и минимальное значение f^* функции $f(x) = x^4 + 8x^3 - 6x^2 - 72x$ на отрезке $[1,5;\ 2]$.

Решение. Функция $f(x)$ на заданном отрезке унимодальна, так как по второму критерию унимодальности $f''(x) = 12x^2 + 48x - 12 \geqslant 0$ на промежутках $x \in (-\infty, -1 - \sqrt{2}] \cup [-1 + \sqrt{2}, \infty)$, где $-1 \pm \sqrt{2}$ — корни квадратного трехчлена $12x^2 + 48x - 12$. Выбрав $n = (2 - 1{,}5)/0{,}05 = 10$, вычислим значение $f(x_i)$ в точках $x_i = 1{,}5 + i \cdot 0{,}05$, $i = \overline{0,\ 10}$, поместив их в таблицу:

x_i	1,50	1,55	1,60	1,65	1,70	1,75
$f(x_i)$	−89,4	−90,2	−91,2	−91,8	−92,08	−92,12

x_i	1,80	1,85	1,90	1,95	2,00
$f(x_i)$	−91,9	−91,4	−90,5	−89,4	−88,0

Путем попарного сравнения значений $f(x_i)$, $i = \overline{0, 10}$, находим $x^* \approx 1{,}75$, $f^* = -92{,}12$.

5.2.2. Метод деления отрезка пополам. Метод деления отрезка пополам сродни итерационному методу деления пополам уточнения корней нелинейных уравнений. Он позволяет построить систему вложенных отрезков:

$$[a_k, b_k] \subseteq [a_{k-1}, b_{k-1}] \subseteq \ldots \subseteq [a, b],$$

внутри которых содержится точка минимума x^*.

Пусть ε — требуемая точность определения x^*. Выбрав $\delta \in (0; 2\varepsilon)$, построим последовательности $\{a_k\}$, $\{b_k\}$, $\left\{x_{\text{лев}}^{(k)}\right\}$,

$\left\{x_{\text{пр}}^{(k)}\right\}$, $k = 0, 1, 2, \ldots$ (из интервала $(0, 2\varepsilon)$ δ выбирается как можно меньше) по следующей схеме (см. рис. 5.2):

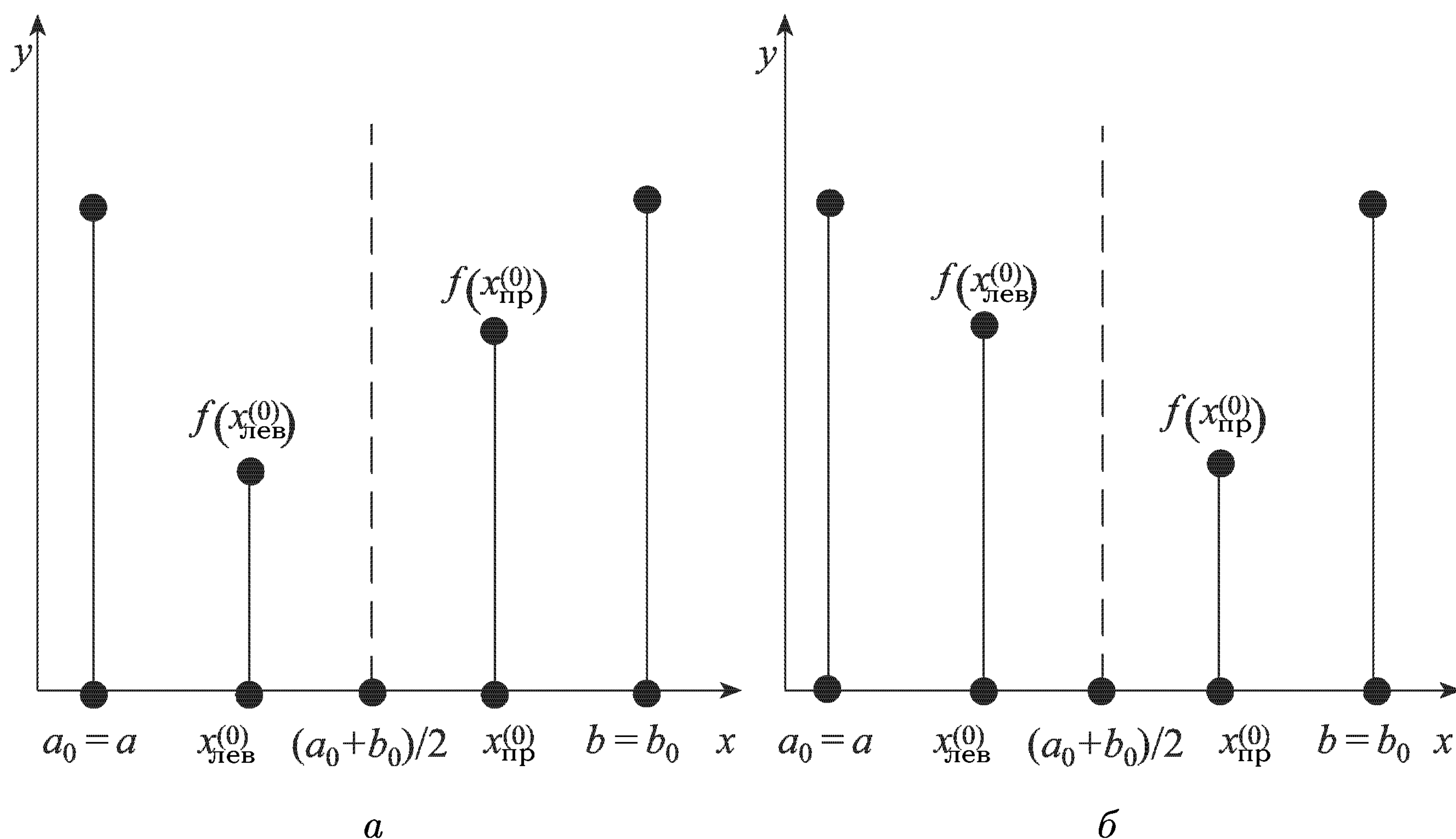


Рис. 5.2. Геометрическая интерпретация метода половинного деления

$$a_0 \equiv a; \quad b_0 \equiv b;$$

$$1)\ \begin{cases} x_{\text{лев}}^{(0)} = \dfrac{a_0 + b_0 - \delta}{2}, \quad x_{\text{пр}}^{(0)} = \dfrac{a_0 + b_0 + \delta}{2}; \\ a_1 = a_0, \quad b_1 = x_{\text{пр}}^{(0)}, \quad \text{если } f\left(x_{\text{лев}}^{(0)}\right) \leqslant f\left(x_{\text{пр}}^{(0)}\right) \ \text{(рис. 5.2 } a\text{);} \\ a_1 = x_{\text{лев}}^{(0)}, \quad b_1 = b_0, \quad \text{если } f\left(x_{\text{лев}}^{(0)}\right) \geqslant f\left(x_{\text{пр}}^{(0)}\right) \ \text{(рис. 5.2 } б\text{);} \end{cases}$$

. . . .

$$k)\ \begin{cases} x_{\text{лев}}^{(k-1)} = \dfrac{a_{k-1} + b_{k-1} - \delta}{2}, \quad x_{\text{пр}}^{(k-1)} = \dfrac{a_{k-1} + b_{k-1} + \delta}{2}; \\ a_k = a_{k-1}, \quad b_k = x_{\text{пр}}^{(k-1)}, \quad \text{если } f\left(x_{\text{лев}}^{(k-1)}\right) \leqslant f\left(x_{\text{пр}}^{(k-1)}\right); \\ a_k = x_{\text{лев}}^{(k-1)}, \quad b_k = b_{k-1}, \quad \text{если } f\left(x_{\text{лев}}^{(k-1)}\right) \geqslant f\left(x_{\text{пр}}^{(k-1)}\right). \end{cases}$$

Процесс останавливается, когда выполняется неравенство $\varepsilon_k = \dfrac{b_k - a_k}{2} \leqslant \varepsilon$.

Окончательное значение точки минимума x^* вычисляется следующим образом: $x^* \approx \dfrac{a_k + b_k}{2}$, $f^* \approx f(x^*)$.

Метод деления отрезка пополам использует значительно меньшее количество вычислений функции $f(x)$, чем метод перебора, т. е. является существенно более экономичным.

Пример 5.3. Методом деления отрезка пополам решить задачу из примера 5.2.

Решение. Положим $\delta = 0{,}02 < 2\varepsilon = 0{,}1$ и построим последовательность $[a_k,\ b_k]$ вложенных отрезков по выше приведенному алгоритму:

$$a_0 = 1{,}5, \quad b_0 = 2{,}0$$

1. $x_{\text{лев}}^{(0)} = \dfrac{a_0 + b_0 - \delta}{2} = \dfrac{3{,}5 - 0{,}02}{2} = 1{,}74$; $x_{\text{пр}}^{(0)} = \dfrac{a_0 + b_0 + \delta}{2} = 1{,}76$; $f\left(x_{\text{лев}}^{(0)}\right) = -92{,}135$; $f\left(x_{\text{пр}}^{(0)}\right) = -92{,}096$. Поскольку $f\left(x_{\text{лев}}^{(0)}\right) < f\left(x_{\text{пр}}^{(0)}\right)$, то $x^* \in \left(a_0, x_{\text{пр}}^{(0)}\right)$; тогда $a_1 = a_0 = 1{,}5$; $b_1 = x_{\text{пр}}^{(0)} = 1{,}76$; $\varepsilon_1 = \dfrac{b_1 - a_1}{2} = \dfrac{1{,}76 - 1{,}5}{2} = 0{,}13 > \varepsilon = 0{,}05$.

2. $x_{\text{лев}}^{(1)} = \dfrac{a_1 + b_1 - \delta}{2} = \dfrac{1{,}5 + 1{,}76 - 0{,}02}{2} = 1{,}62$; $x_{\text{пр}}^{(1)} = \dfrac{a_1 + b_1 + \delta}{2} = 1{,}64$; $f\left(x_{\text{лев}}^{(1)}\right) = -91{,}486$; $f\left(x_{\text{пр}}^{(1)}\right) = -91{,}696$. Поскольку $f\left(x_{\text{лев}}^{(1)}\right) > f\left(x_{\text{пр}}^{(1)}\right)$, то $x^* \in \left(x_{\text{лев}}^{(1)}, b_1\right)$; тогда $a_2 = x_{\text{лев}}^{(1)} = 1{,}62$; $b_2 = b_1 = 1{,}76$; $\varepsilon_2 = \dfrac{b_2 - a_2}{2} = \dfrac{1{,}76 - 1{,}62}{2} = 0{,}07 > \varepsilon = 0{,}05$.

3. $x_{\text{лев}}^{(2)} = \dfrac{a_2 + b_2 - \delta}{2} = \dfrac{1{,}62 + 1{,}76 - 0{,}02}{2} = 1{,}68$; $x_{\text{пр}}^{(2)} = \dfrac{a_2 + b_2 + \delta}{2} = 1{,}70$; $f\left(x_{\text{лев}}^{(2)}\right) = -91{,}995$; $f\left(x_{\text{пр}}^{(2)}\right) = -92{,}084$. Поскольку $f\left(x_{\text{лев}}^{(2)}\right) > f\left(x_{\text{пр}}^{(2)}\right)$, то $x^* \in \left(x_{\text{лев}}^{(2)}, b_2\right)$; тогда $a_3 = x_{\text{лев}}^{(2)} = 1{,}68$; $b_3 = b_2 = 1{,}76$; $\varepsilon_3 = \dfrac{b_3 - a_3}{2} = \dfrac{1{,}76 - 1{,}68}{2} = 0{,}04 < \varepsilon = 0{,}05$.

Заданная точность достигнута, и, следовательно $x^* = \dfrac{a_3 + b_3}{2} = 1{,}72$; $f^* = f\,(1{,}72) = -92{,}13$.

5.2.3. Метод золотого сечения. Метод золотого сечения так же является *прямым* методом, не использующим производные функции $f(x)$. Он позволяет сократить вычисления функции $f(x)$ по сравнению с методом деления пополам, вычисляя не *два значения* на каждой итерации, а *одно*.

Деление отрезка на две неравные части так, что отношение длины всего отрезка к длине большей части равно отношению длины большей части к длине меньшей части, называется *золотым сечением.*

Золотое сечение отрезка $[a, b]$ осуществим двумя точками x_1, x_2 (рис. 5.3):

$$x_1 = a + \frac{3-\sqrt{5}}{2}(b-a), \quad x_2 = a + \frac{\sqrt{5}-1}{2}(b-a),$$

причем $x_1 < x_2$ и x_1 — вторая (большая) точка золотого сечения отрезка $[a, x_2]$, а x_2 — первая (меньшая) точка золотого сечения отрезка $[x_1, b]$. Эти значения x_1 и x_2 получены следующим образом.

Пусть $x_1 - a = l_1$; $b - x_1 = l_2$, $l = l_1 + l_2 = b - a$. Тогда по определению золотого сечения

$$\frac{l}{l_2} = \frac{l_2}{l_1},$$

откуда

$$l_2^2 = (l_1 + l_2)l_1; \quad \left(\frac{l_1}{l_2}\right)^2 + \frac{l_1}{l_2} - 1 = 0; \quad \frac{l_1}{l_2} = \frac{\sqrt{5}-1}{2};$$

$$\begin{cases} l_1 = \dfrac{\sqrt{5}-1}{2} l_2, \\ l_1 + l_2 = l. \end{cases}$$

Из этой системы находим

$$l_1 = \frac{3-\sqrt{5}}{2} l; \quad l_2 = \frac{\sqrt{5}-1}{2} l.$$

Отсюда получаем

$$\begin{cases} x_1 = a + \dfrac{3-\sqrt{5}}{2}(b-a) = a + 0{,}38(b-a), \\ x_2 = a + \dfrac{\sqrt{5}-1}{2}(b-a) = a + 0{,}62(b-a), \end{cases}$$

причем, зная одну из точек золотого сечения, можно найти другую из равенств

$$x_1 = a + b - x_2, \quad x_2 = a + b - x_1.$$

Таким образом, алгоритм метода золотого сечения следующий (см. рис. 5.3 *а* и *б*):

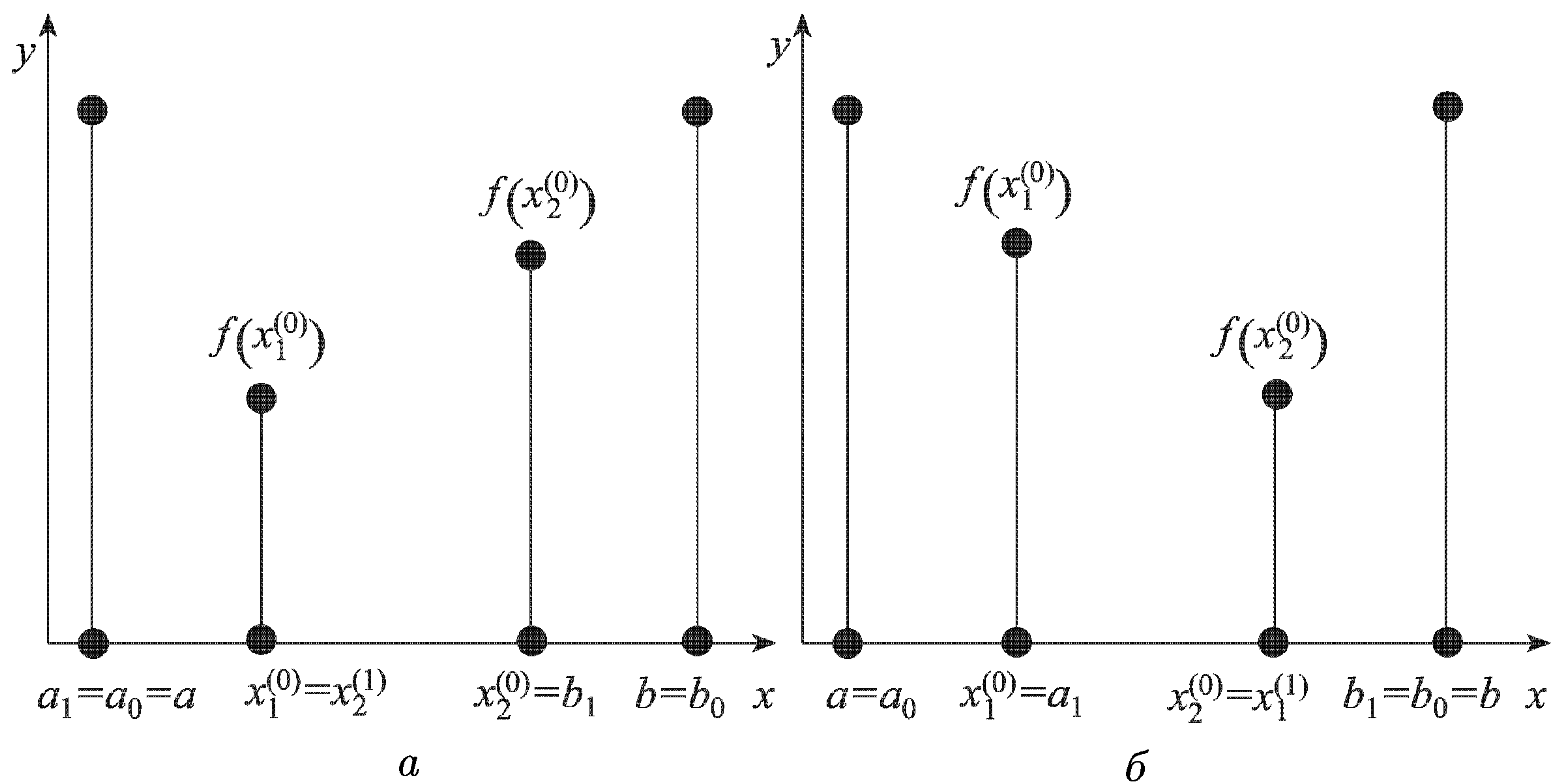


Рис. 5.3. К методу золотого сечения

Строится система вложенных отрезков

$$[a_k, b_k] \subseteq [a_{k-1}, b_{k-1}] \subseteq \ldots \subseteq [a, b],$$

внутри которых находится точка минимума x^*, и последовательности золотых сечений $x_1^{(0)}$, $x_2^{(0)}$; $x_1^{(1)}$, $x_2^{(1)}$; $\ldots$; $x_1^{(k)}$, $x_2^{(k)}$; $\ldots$, причем $x_1^{(k)} < x_2^{(k)}$, $k = 0, 1, 2, \ldots$, тогда

$$a_0 \equiv a; \quad b_0 \equiv b;$$

Рис. 5.3 *а* $\begin{cases} a_1 = a_0;\ b_1 = x_2^{(0)}; \\ x_2^{(1)} = x_1^{(0)},\ \text{ если }\ f(x_1^{(0)}) \leqslant f(x_2^{(0)}), \\ \text{т. е. вторая точка } x_2^{(1)} \text{золотого сечения на} \\ \text{следующем шаге имеется, вычисляем первую } x_1^{(1)}: \\ x_1^{(1)} = a_1 + b_1 - x_2^{(1)}, \quad x^* \in (a_0, x_2^{(0)}) \equiv (a_1, b_1); \end{cases}$

Рис. 5.3 *б* $\begin{cases} a_1 = x_1^{(0)};\ \ b_1 = b_0; x_1^{(1)} = x_2^{(0)},\ \text{ если }\ f(x_1^{(0)}) > f(x_2^{(0)}), \\ \text{т. е. первая точка } x_1^{(1)} \text{ золотого сечения на} \\ \text{следующем шаге имеется, вычисляем вторую } x_2^{(1)}: \\ x_2^{(1)} = a_1 + b_1 - x_1^{(1)}, \quad x^* \in (x_1^{(0)}, b_0) \equiv (a_1, b_1); \end{cases}$

. . . .

Рис. 5.3 *а* $\begin{cases} a_k = a_{k-1}; \quad b_k = x_2^{(k-1)}; \\ x_2^{(k)} = x_1^{(k-1)}, \quad \text{если} \quad f(x_1^{(k-1)}) \leqslant f(x_2^{(k-1)}), \\ x_1^{(k)} = a_k + b_k - x_2^{(k)}, \quad x^* \in (a_{k-1}, x_2^{(k-1)}) \equiv (a_k, b_k); \end{cases}$

Рис. 5.3 *б* $\begin{cases} a_k = x_1^{(k-1)}; \quad b_k = b_{k-1}; \\ x_1^{(k)} = x_2^{(k-1)}, \quad \text{если} \quad f(x_1^{(k-1)}) > f(x_2^{(k-1)}), \\ x_2^{(k)} = a_k + b_k - x_1^{(k)}, \quad x^* \in (x_1^{(k-1)}, b_{k-1}) \equiv (a_k, b_k). \end{cases}$

Значение x^* определяется следующим образом:

$$x^* \approx \overline{x} = \begin{cases} x_1^{(k)}, \text{ если } f(x_1^{(k)}) \leqslant f(x_2^{(k)}), \\ x_2^{(k)}, \text{ если } f(x_1^{(k)}) > f(x_2^{(k)}). \end{cases}$$

Останов происходит тогда, когда выполняется неравенство $|b_k - a_k| \leqslant \varepsilon$.

Погрешность метода золотого сечения оценивается по формуле

$$|x^* - \overline{x}| \leqslant \left(\frac{\sqrt{5} - 1}{2}\right)^k (b - a),$$

и если задана точность ε, то, поскольку погрешность не превышает точности, из неравенства

$$\left(\frac{\sqrt{5} - 1}{2}\right)^k (b - a) \leqslant \varepsilon$$

до начала процесса вычислений находим нижнюю оценку числа шагов k:

$$k \geqslant \ln \frac{\varepsilon}{b - a} \Big/ \ln \left(\frac{\sqrt{5} - 1}{2}\right) \approx -2{,}1 \cdot \ln \frac{\varepsilon}{b - a}.$$

Пример 5.4. Решить задачу примера 5.2 методом золотого сечения.

Решение. В соответствии с алгоритмом $a_0 = 1{,}5$, $b_0 = 2$; $x_1^{(0)} = a_0 + \dfrac{3 - \sqrt{5}}{2}(b_0 - a_0) = 1{,}691$; $x_2^{(0)} = a_0 + \dfrac{\sqrt{5} - 1}{2}(b_0 - a_0) = 1{,}809$.

1-й шаг. $f\left(x_1^{(0)}\right) = -92{,}049$; $f\left(x_2^{(0)}\right) = -91{,}814$. Поскольку $f\left(x_1^{(0)}\right) < f\left(x_2^{(0)}\right)$, то $x^* \in \left(a_0,\ x_2^{(0)}\right)$; $a_1 = a_0 = 1{,}5$; $b_1 = x_2^{(0)} = 1{,}809$; $\varepsilon_1 = b_1 - a_1 = 1{,}809 - 1{,}5 = 0{,}309 > \varepsilon = 0{,}05$; $x_2^{(1)} = x_1^{(0)} = 1{,}691$.

2-й шаг. Значение $x_2^{(1)} = 1{,}691$ имеется, а также имеется $f\left(x_2^{(1)}\right) = f\left(x_1^{(0)}\right) = -92{,}049$. Вычисляем $x_1^{(1)} = a_1 + b_1 - x_2^{(1)} = 1{,}5 + 1{,}809 - 1{,}691 = 1{,}618$ и $f\left(x_1^1\right) = f\,(1{,}618) = -91{,}464$. Поскольку $f\left(x_1^{(1)}\right) > f\left(x_2^{(1)}\right)$, то $x^* \in \left(x_1^{(1)}, b_1\right) = (1{,}618;\ 1{,}809)$; $a_2 = x_1^{(1)} = 1{,}618$; $b_2 = b_1 = 1{,}809$; $\varepsilon_2 = b_2 - a_2 = 1{,}809 - 1{,}618 = 0{,}191 > \varepsilon = 0{,}05$; $x_1^{(2)} = x_2^{(1)} = 1{,}691$.

3-й шаг. Значения $x_1^{(2)} = 1{,}691$ и $f\left(x_1^{(2)}\right) = f\left(x_2^{(1)}\right) = -92{,}049$ уже имеются. Вычисляем $x_2^{(2)} = a_2 + b_2 - x_1^{(2)} = 1{,}618 + 1{,}809 - 1{,}691 = 1{,}736$ и $f\left(x_2^{(2)}\right) = f\,(1{,}736) = -92{,}138$. Поскольку $f\left(x_1^{(2)}\right) > f\left(x_2^{(2)}\right)$, то $x^* \in \left(x_1^{(2)};\ b_2\right) = (1{,}691;\ 1{,}809)$; $a_3 = x_1^{(2)} = 1{,}691$; $b_3 = b_2 = 1{,}809$; $\varepsilon_3 = b_3 - a_3 = 1{,}809 - 1{,}691 = 0{,}118 > \varepsilon = 0{,}05$; $x_1^{(3)} = x_2^{(2)} = 1{,}736$.

4-й шаг. Значения $x_1^{(3)} = 1{,}736$ и $f\left(x_1^{(3)}\right) = f\left(x_2^{(2)}\right) = -92{,}138$ уже имеются. Вычисляем $x_2^{(3)} = a_3 + b_3 - x_1^3 = 1{,}691 + 1{,}809 - 1{,}736 = 1{,}764$ и $f\left(x_2^{(3)}\right) = f\,(1{,}764) = -92{,}083$. Поскольку $f\left(x_1^{(3)}\right) < f\left(x_2^{(3)}\right)$, то $x^* \in \left(a_3, x_2^{(3)}\right) = (1{,}691;\ 1{,}764)$; $a_4 = a_3 = 1{,}691$; $b_4 = x_2^{(3)} = 1{,}764$; $\varepsilon_4 = b_4 - a_4 = 1{,}764 - 1{,}691 = 0{,}073 > \varepsilon = 0{,}05$; $x_2^{(4)} = x_1^{(3)} = 1{,}736$.

5-й шаг. Значения $x_2^{(4)} = 1{,}736$ и $f\left(x_2^{(4)}\right) = f\left(x_1^{(3)}\right) = -92{,}138$ уже имеются; $x_1^{(4)} = a_4 + b_4 - x_2^{(4)} = 1{,}691 + 1{,}764 - 1{,}736 = 1{,}719$; $f\left(x_1^{(4)}\right) = f\,(1{,}719) = -92{,}129$. Поскольку $f\left(x_1^{(4)}\right) > f\left(x_2^{(4)}\right)$, то $x^* \in \left(x_1^{(4)},\ b_4\right) = (1{,}719;\ 1{,}764)$; $a_5 = 1{,}719$; $b_5 = 1{,}764$; $\varepsilon_5 = b_5 - a_5 = 1{,}764 - 1{,}719 = 0{,}045 < \varepsilon = 0{,}05$.

Требуемая точность достигнута. Поскольку $f\left(x_2^{(4)}\right) = -92{,}138 < f\left(x_1^{(4)}\right) = -92{,}129$, то в качестве точки минимума принимается точка $x^* \approx x_2^{(4)} = 1{,}736$; $f^* = -92{,}138$.

§ 5.3. Методы минимизации, использующие производные. Метод Ньютона

Метод Ньютона, как метод с квадратичной скоростью сходимости, используется на завершающем этапе какого-либо более грубого метода.

Он использует производные 1-го и 2-го порядков, что обеспечивает очень быструю сходимость.

Множество точек $X \subset R$ называется *выпуклым*, если отрезок $[x_1, x_2]$, соединяющий любые две точки x_1, $x_2 \in X$, принадлежит множеству X, т. е. точки $\alpha\, x_1 + (1-\alpha)x_2 \in X$, $0 < \alpha < 1$.

Функция $f(x)$ называется *выпуклой* на множестве точек $x \in \in X$, если для двух точек x_1, $x_2 \in X$ выполняется неравенство

$$f(\alpha\, x_1 + (1-\alpha)x_2) \leqslant \alpha\, f(x_1) + (1-\alpha)f(x_2), \quad 0 < \alpha < 1.$$

Будем рассматривать *выпуклые функции* $f(x)$ на отрезке $x \in \in [a, b]$.

Для того чтобы *дважды дифференцируемая* функция была *выпуклой* на отрезке $x \in [a, b]$, необходимо и достаточно, чтобы выполнялось неравенство $f''(x) \geqslant 0$ для всех точек.

Для выпуклой дважды дифференцируемой функции $f(x)$ на $x \in [a, b]$ *метод Ньютона* заключается в построении следующей итерационной последовательности (сравнить с методом Ньютона уточнения корней нелинейных уравнений (2.46), (2.47)):

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}, \qquad k = 0, 1, 2\ldots$$

Отсюда видно, что производная второго порядка $f''(x)$ на отрезке $x \in [a, b]$ не должна быть равна нулю. Для выпуклых функций $f(x)$ это действительно так, в противном случае точки, в которых $f''(x) = 0$, являлись бы точками перегиба функции, а в этих точках минимум (или максимум) функции $f(x)$ отсутствует.

Если задана точность ε, то останов осуществляется при выполнении условия

$$f'(x_k) \leqslant \varepsilon,$$

т. е. когда касательная к графику функции в точке $(x_k,\ f(x_k))$ почти горизонтальна. Тогда

$$x^* \approx x_k \ \text{ и } \ f(x^*) \approx f(x_k).$$

Можно показать, что верхняя оценка погрешности в методе Ньютона представима в виде следующего неравенства:

$$|x^* - x_k| \leqslant \frac{2m_2}{M_1} q^{2^k}, \quad q = \frac{M_1}{2m_2} \left|f'(x_0)\right|,$$

$$k = 0,\, 1,\, 2,\, \ldots\, ; \quad M_1 = \max_{x \in [a,\, b]} \left|f'(x)\right|; \quad m_2 = \min_{x \in [a, b]} \left|f''(x)\right|,$$

причем для сходимости метода Ньютона достаточно, чтобы начальное приближение x_0 удовлетворяло условию $q < 1$.

Пример 5.5. Методом Ньютона найти точку минимума x^* и минимальное значение f^* функции $f(x) = (x-2)^4 - \ln x$ на отрезке $x \in [2; 3]$ с точностью $\varepsilon = |f'(x_k)| < 10^{-7}$.

Решение. В окрестности этой точки функция $f(x)$ является выпуклой, поскольку $f''(x) = 12(x-2)^2 + \dfrac{1}{x^2} > 0$. Вычисления по методу Ньютона сводятся в следующую таблицу (за x_0 принимается любой конец отрезка [2;3]):

k	x_k	$f(x_k)$	$f'(x_k)$
0	3	$-9{,}86 \cdot 10^{-2}$	3,67
1	2,6972477	−0,7558859	0,985
2	2,5322701	−0,8488508	0,208
3	2,4736906	−0,8553636	0,0201
4	2,4663735	−0,8554408	0,0003
5	2,4662656	−0,8554408	$5 \cdot 10^{-8} < 10^{-7}$

Окончательно получаем $x^* \approx 2{,}4662656$, $f^* \approx -0{,}8554408$.

§ 5.4. Безусловная минимизация функций многих переменных

5.4.1. Метод градиентного спуска. Будем рассматривать выпуклые функции многих переменных $f(x)$, $x = (x_1\, x_2\, \ldots\, x_n)^T$, определенные на выпуклых множествах X из n-мерного евклидова пространства $x \in X \subset E^n$. Для того чтобы функция одной переменной была выпуклой, необходимо и достаточно, чтобы 2-ая производная на всем отрезке $[a,\ b]$ была неотрицательна.

Для функций многих переменных используется следующий критерий выпуклости.

Если $f(x)$ — дважды дифференцируемая на выпуклом множестве $X \subset E^n$ функция и матрица ее частных производных второго порядка (матрица Гессе) $\left[\dfrac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]$, $i, j = \overline{1, n}$, положительно определена для всех $x \in X$, то функция $f(x)$ является выпуклой на множестве точек X.

С использованием критерия Сильвестра формулировка этого критерия упрощается и принимает следующий вид.

Если все диагональные миноры матрицы $\left[f''_{x_i x_j}(x)\right]$, $i, j = \overline{1, n}$, положительны на множестве точек $x \in X$, то функция $f(x)$ выпукла на этом множестве точек X.

Понятие выпуклости функций является исключительно важным, поскольку в численных методах оптимизации выпуклых на множестве $x \in X$ функций задача о нахождении локального минимума на всем множестве точек $x \in X$ совпадает с задачей нахождения глобального минимума на этом множестве.

Пример 5.6. Выяснить, является ли функция $f(x_1,\ x_2) = 2x_1^2 + x_2^2 + \sin(x_1 + x_2)$ выпуклой в двумерном пространстве E^2.

Решение. Составим матрицу вторых производных:

$$f''(x) = \begin{bmatrix} f''_{x_1 x_1} & f''_{x_1 x_2} \\ f''_{x_2 x_1} & f''_{x_2 x_2} \end{bmatrix} = \begin{bmatrix} 4 - \sin(x_1 + x_2) & -\sin(x_1 + x_2) \\ -\sin(x_1 + x_2) & 2 - \sin(x_1 + x_2) \end{bmatrix},$$

из которой видно, что диагональный минор первого порядка $\Delta_1 = 4 - \sin(x_1 + x_2) \geqslant 3 > 0$ на всей плоскости $x_1 O x_2$, поскольку $|\sin(x_1 + x_2)| \leqslant 1$. Минор второго порядка $\Delta_2 = 8 -$

$-6\sin(x_1+x_2) \geqslant 2 > 0$ по той же причине. На основании критерия Сильвестра заключаем, что матрица вторых производных $f''(x)$ положительно определена на всей плоскости x_1Ox_2, а следовательно, заданная функция является выпуклой на этом множестве.

Пусть $f(x)$, $x = (x_1 \ldots x_n)^T$, — выпуклая дифференцируемая функция на всем множестве точек $x \in X$ в евклидовом пространстве E^n. Требуется найти точку ее минимума x^* и минимум $f(x^*)$. Выбрав произвольное начальное приближение $x^{(0)} \in \in X \subset E^n$, построим следующую итерационную последовательность:

$$x^{(k+1)} = x^{(k)} - \alpha_k \operatorname{grad} f(x^{(k)}), \quad k = 0, 1, 2 \ldots, \tag{5.1}$$

где величины α_k (параметрические шаги) выбираются *достаточно малыми* из условия

$$f(x^{(k+1)}) < f(x^{(k)}), \quad k = 0, 1, 2 \ldots \tag{5.2}$$

Остановимся подробнее на векторном соотношении (5.1). Поскольку градиент функции многих переменных в ортонормированном базисе $e_1, e_2, \ldots, e_n$ определяется как вектор $\operatorname{grad} f(x) = = \dfrac{\partial f(x)}{\partial x_1}e_1 + \dfrac{\partial f(x)}{\partial x_2}e_2 + \ldots + \dfrac{\partial f(x)}{\partial x_n}e_n$, то координатами вектора градиента функции многих переменных являются частные производные $\dfrac{\partial f(x)}{\partial x_i}$, $i = \overline{1, n}$, этой функции по переменным x_i, $i = = \overline{1, n}$.

Тогда алгоритм (5.1) в скалярной форме принимает вид

$$\begin{aligned}
x_1^{(k+1)} &= x_1^{(k)} - \alpha_k \frac{\partial f\left(x^{(k)}\right)}{\partial x_1},\\
x_2^{(k+1)} &= x_2^{(k)} - \alpha_k \frac{\partial f\left(x^{(k)}\right)}{\partial x_2},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
x_n^{(k+1)} &= x_n^{(k)} - \alpha_k \frac{\partial f\left(x^{(k)}\right)}{\partial x_n}.
\end{aligned}$$

Окончание процесса устанавливается по близости к нулю $\mathrm{grad}\, f(x^{(k)})$, т. е. при выполнении неравенств

$$\left| \frac{\partial f\left(x^{(k)}\right)}{\partial x_i} \right| \leqslant \varepsilon, \quad i = \overline{1, n}, \tag{5.3}$$

где ε — заданная точность, или

$$\left|\mathrm{grad}\, f\left(x^{(k)}\right)\right| = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial f\left(x^{(k)}\right)}{\partial x_i} \right)^2} \leqslant \varepsilon. \tag{5.4}$$

Если условие (5.2) не выполняется, то α_k уменьшается вдвое и алгоритм (5.1) повторяется и т. д. При выполнении условий (5.3) или (5.4) полагают

$$x^* \approx x^{(k)} \text{ и } f(x^*) \approx f(x^{(k)}).$$

Алгоритм (5.1) для функции $f(x, y)$ двух переменных приобретает вид

$$\begin{cases} x^{(k+1)} = x^{(k)} - \alpha_k \dfrac{\partial f(x^{(k)}, y^{(k)})}{\partial x}, \\ y^{(k+1)} = y^{(k)} - \alpha_k \dfrac{\partial f(x^{(k)}, y^{(k)})}{\partial y}. \end{cases}$$

Пример 5.7. Методом градиентного спуска с точностью $\varepsilon = 0{,}05$ минимизировать функцию $f(x) = f(x_1, x_2) = x_1^2 + 2x_2^2 + \exp(x_1 + x_2)$.

Решение. Вначале необходимо проверить выпуклость функции на множестве $x \in X \subset E^2$, для чего составляется матрица частных производных второго порядка:

$$f''(x) = \begin{bmatrix} 2 + \exp(x_1 + x_2) & \exp(x_1 + x_2) \\ \exp(x_1 + x_2) & 4 + \exp(x_1 + x_2) \end{bmatrix} = 8 + 6\exp(x_1 + x_2) > 0,$$

т. е.

$$\Delta_2 > 0; \quad \Delta_1 = 2 + \exp(x_1 + x_2) > 0.$$

Таким образом, по критерию Сильвестра заданная функция выпукла на всей плоскости $x_1 O x_2$.

Для решения задачи выберем начальное приближение $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}\right) = (0;\ 0)$ и $\alpha_0 = 1$.

1-й шаг: $k = 0$; $x_1^{(0)} = 0$; $x_2^{(0)} = 0$; $\alpha_0 = 1$; $f\left(x_1^{(0)}, x_2^{(0)}\right) = 1$, $\dfrac{\partial f\left(x^{(0)}\right)}{\partial x_1} = 1$; $\dfrac{\partial f\left(x^{(0)}\right)}{\partial x_2} = 1$. По формуле (5.1)

$$x_1^{(1)} = x_1^{(0)} - \alpha_0 \frac{\partial f\left(x^{(0)}\right)}{\partial x_1} = -1; \quad x_2^{(1)} = x_2^{(0)} - \alpha_0 \frac{\partial f\left(x^{(0)}\right)}{\partial x_2} = -1;$$

$$f\left(x_1^{(1)},\ x_2^{(1)}\right) = 3{,}145 > f\left(x_1^{(0)},\ x_2^{(0)}\right) = 1,$$

т. е. условие (5.2) не выполнено, вследствие чего необходимо уменьшить параметрический шаг α_0. Уменьшаем его вдвое, приняв $\alpha_0 = 0{,}5$, и повторяем по (5.1) вычисления с $x_1^{(0)} = 0$ и $x_2^{(0)} = 0$:

$$x_1^1 = 0 - 0{,}5 \cdot 1 = -0{,}5; \quad x_2^1 = 0 - 0{,}5 \cdot 1 = -0{,}5;$$

$$f\left(x_1^{(1)},\ x_2^{(1)}\right) = 1{,}118 > f\left(x_1^{(0)},\ x_1^{(0)}\right) = 1;$$

условие (5.2) не выполнено. Уменьшаем α_0 вдвое: $\alpha_0 = 0{,}25$:

$$x_1^{(1)} = x_1^{(0)} - \alpha_0 \frac{\partial f\left(x^{(0)}\right)}{\partial x_1} = 0 - 0{,}25 \cdot 1 = -0{,}25;$$

$$x_2^{(1)} = x_2^{(0)} - \alpha_0 \frac{\partial f\left(x^{(0)}\right)}{\partial x_2} = 0 - 0{,}25 \cdot 1 = -0{,}25;$$

$$f\left(x_1^{(1)},\ x_2^{(1)}\right) = f\,(-0{,}25;\ -0{,}25) = 0{,}794 < f\left(x_1^{(0)},\ x_2^{(0)}\right) = 1;$$

$$\frac{\partial f\left(x_1^{(1)},\ x_2^{(1)}\right)}{\partial x_1} = 2x_1^{(1)} + \exp\left(x_1^{(1)} + x_2^{(1)}\right) = \\ = -0{,}5 + \exp\,(-0{,}25 - 0{,}25) = 0{,}106;$$

$$\frac{\partial f\left(x_1^{(1)},\ x_2^{(1)}\right)}{\partial x_2} = 4x_2^{(1)} + \exp\left(x_1^{(1)} + x_2^{(1)}\right) = \\ = 4\,(-0{,}25) + \exp\left(x_1^{(1)} + x_2^{(1)}\right) = -0{,}393;$$

$$\left|\operatorname{grad} f\left(x^{(1)}\right)\right| = \sqrt{0{,}106^2 + (-0{,}393)^2} = 0{,}407 > \varepsilon = 0{,}05.$$

2-й шаг: $k = 1$; $x_1^{(1)} = -0{,}25$; $x_2^{(1)} = -0{,}25$; $\alpha_1 = 0{,}25$;

$$x_1^{(2)} = x_1^{(1)} - \alpha_1 \frac{\partial f\left(x^{(1)}\right)}{\partial x_1} = -0{,}25 - 0{,}25 \cdot 0{,}106 = -0{,}2765;$$

$$x_2^{(2)} = x_2^{(1)} - \alpha_1 \frac{\partial f\left(x^{(1)}\right)}{\partial x_2} = -0{,}25 - 0{,}25 \cdot (-0{,}393) = -0{,}1518;$$

$$f\left(x_1^{(2)}, x_2^{(2)}\right) = 0{,}774 < f\left(x_1^{(1)}, x_2^{(1)}\right) = 0{,}794.$$

Сохраняем значение α_1, приняв $\alpha_2 = 0{,}25$:

$$\frac{\partial f\left(x_1^{(2)},\, x_2^{(2)}\right)}{\partial x_1} = 2x_1^{(2)} + \exp\left(x_1^{(2)} + x_2^{(2)}\right) = 0{,}0983;$$

$$\frac{\partial f\left(x_1^{(2)},\, x_2^{(2)}\right)}{\partial x_2} = 4x_2^{(2)} + \exp\left(x_1^{(2)} + x_2^{(2)}\right) = 0{,}0451;$$

$$\left|\operatorname{grad} f\left(x^{(2)}\right)\right| = \sqrt{0{,}0983^2 + 0{,}0451^2} = 0{,}108 > \varepsilon = 0{,}05.$$

3-й шаг: $k = 2$; $x_1^{(2)} = -0{,}2765$; $x_2^{(2)} = -0{,}1518$; $\alpha_2 = 0{,}25$;

$$x_1^{(3)} = x_1^{(2)} - \alpha_2 \frac{\partial f\left(x^{(2)}\right)}{\partial x_1} = -0{,}2765 - 0{,}25 \cdot 0{,}0983 = -0{,}301;$$

$$x_2^{(3)} = x_2^{(2)} - \alpha_2 \frac{\partial f\left(x^{(2)}\right)}{\partial x_2} = -0{,}1518 - 0{,}25 \cdot 0{,}0451 = -0{,}163;$$

$$f\left(x_1^{(3)},\, x_2^{(3)}\right) = 0{,}772 < 0{,}774;$$

$$\left|\operatorname{grad}\ f\left(x^{(3)}\right)\right| = \sqrt{\left(\frac{\partial f\left(x^{(3)}\right)}{\partial x_1}\right)^2 + \left(\frac{\partial f\left(x^{(3)}\right)}{\partial x_2}\right)^2} =$$

$$= \sqrt{0{,}0262^2 + (-0{,}023)^2} = 0{,}03486 < \varepsilon.$$

Точность достигнута, следовательно, $x^* = (x_1^*;\, x_2^*) \approx \approx (-0{,}301;\, -0{,}163)$; $f^* \approx 0{,}772$.

5.4.2. Метод наискорейшего спуска. Метод наискорейшего спуска отличается от метода градиентного спуска (5.1) *оптимальным способом определения параметрического шага* α_k, который находится из условия

$$\psi_k(\alpha_k) = \min_{\alpha>0} \psi_k(\alpha), \tag{5.5}$$

где

$$\psi_k(\alpha) = f(x^{(k)} - \alpha \operatorname{grad} f(x^{(k)})). \tag{5.6}$$

Здесь

$$f(x^{(k)} - \alpha \operatorname{grad} f(x^{(k)})) =$$

$$= f\left(x_1^{(k)} - \alpha \frac{\partial f(x_1^{\{k\}}, \ldots, x_n^{(k)})}{\partial x_1}, \ldots, x_n^{(k)} - \alpha \frac{\partial f(x_1^{(k)}, \ldots, x_n^{(k)})}{\partial x_n}\right).$$

После минимизации функции $\psi(\alpha)$ одной переменной α найденное значение α^* принимается за α_k, после чего реализуется метод градиентного спуска:

$$x^{(k+1)} = x^{(k)} - \alpha_k \operatorname{grad}\, f(x^{(k)}), \quad k = 0, 1, 2, \ldots$$

Окончание итерационного процесса устанавливается при выполнении условия (5.3) или (5.4).

Таким образом, на каждом шаге метода наискорейшего спуска решается задача минимизации (5.5), (5.6) функции $\psi(\alpha)$ одной переменной α.

При таком выборе α_k достигается максимально возможное уменьшение функции $f(x)$.

Можно показать, что в методе наискорейшего спуска $\operatorname{grad} f(x^{(k)})$ и $\operatorname{grad} f(x^{(k+1)})$ ортогональны, т. е. скалярные произведения $(\operatorname{grad} f(x^{(k)}), \operatorname{grad} f(x^{(k+1)})) = 0$.

Пример 5.8. Методом наискорейшего спуска решить задачу примера 5.7.

Решение.

1-й шаг: $k = 0$.

Выберем начальное приближение

$$x^{(0)} = \left(x_1^{(0)},\, x_2^{(0)}\right) = (0;\, 0)\,.$$

Тогда

$$\frac{\partial f\left(x^{(0)}\right)}{\partial x_1} = 1; \quad \frac{\partial f\left(x^{(0)}\right)}{\partial x_2} = 1;$$

$$\psi_0(\alpha) = f(0-\alpha\cdot 1,\ 0-\alpha\cdot 1) = f(-\alpha,\ -\alpha) = 3\alpha^2 + \exp(2\alpha)\,.$$

Для минимизации $\psi_0(\alpha)$ используем метод перебора, для чего составляется сеточная функция для функции $\psi_0(\alpha)$:

α	0,18	0,20	0,22	0,24	0,26
$\psi_0(\alpha)$	0,7949	0,7903	0,7892	0,7916	0,7973

Отсюда ясно, что $\alpha = \alpha_0 = 0{,}22$ и

$$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \alpha_0 \begin{pmatrix} \partial f\left(x^{(0)}\right)/\partial x_1 \\ \partial f\left(x^{(0)}\right)/\partial x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \\ - 0{,}22\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -0{,}22 \\ -0{,}22 \end{pmatrix};$$

$$\left|\operatorname{grad} f\left(x^{(1)}\right)\right| = \sqrt{\left(\frac{\partial f\left(x^{(1)}\right)}{\partial x_1}\right)^2 + \left(\frac{\partial f\left(x^{(1)}\right)}{\partial x_2}\right)^2} = \\ = \sqrt{(0{,}204)^2 + (-0{,}236)^2} = 0{,}312 > \varepsilon\,.$$

2-й шаг: $k = 1$; $x^{(1)} = \left(x_1^{(1)},\ x_2^{(1)}\right) = (-0{,}22;\ -0{,}22)$;

$$\frac{\partial f\left(x^{(1)}\right)}{\partial x_1} = 0{,}204; \quad \frac{\partial f\left(x^{(1)}\right)}{\partial x_2} = -0{,}236;$$

$$\psi_1(\alpha) = f(-0{,}22 - 0{,}204\alpha;\ -0{,}22 + 0{,}236\alpha) = \\ = (-0{,}22 - 0{,}204\alpha)^2 + \\ + 2(-0{,}22 + 0{,}236\alpha)^2 + \exp(-0{,}44 + 0{,}032\alpha) = \\ = 0{,}1452 - 0{,}1179\alpha + 0{,}153\alpha^2 + \exp(-0{,}44 + 0{,}032\alpha)\,.$$

Минимизируем $\psi_1(\alpha)$:

α	0,28	0,30	0,32	0,34	0,36
$\psi_1(\alpha)$	0,774	0,77384	0,7738	0,77387	0,7741

т. е. $\alpha = \alpha_1 = 0{,}32$;

$$\begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} = \begin{pmatrix} -0{,}22 \\ -0{,}22 \end{pmatrix} - 0{,}32 \begin{pmatrix} 0{,}204 \\ -0{,}236 \end{pmatrix} = \begin{pmatrix} -0{,}2853 \\ -0{,}1445 \end{pmatrix};$$

$$\left|\operatorname{grad} f\left(x^{(2)}\right)\right| = \sqrt{\left(\frac{\partial f\left(x^{(2)}\right)}{\partial x_1}\right)^2 + \left(\frac{\partial f\left(x^{(2)}\right)}{\partial x_2}\right)^2} =$$

$$= \sqrt{0{,}08^2 + 0{,}07264^2} = 0{,}1081 > \varepsilon.$$

3-й шаг: $k = 2$; $x^{(2)} = \left(x_1^{(2)}, x_2^{(2)}\right) = (-0{,}2853;\ -0{,}1445)$;

$$\frac{\partial f\left(x^{(2)}\right)}{\partial x_1} = 0{,}08; \quad \frac{\partial f\left(x^{(2)}\right)}{\partial x_2} = 0{,}07264;$$

$$\psi_2(\alpha) = (-0{,}2853 - 0{,}08\alpha)^2 + \\ + 2(-0{,}1445 - 0{,}07264\alpha)^2 + \exp(-0{,}43 - 0{,}1526\alpha).$$

Минимизируем $\psi_2(\alpha)$:

α	0,20	0,22	0,24	0,26	0,28
$\psi_2(\alpha)$	0,77273	0,77241	0,77240	0,77241	0,77244

т. е. $\alpha = \alpha_2 = 0{,}24$. Тогда

$$\begin{pmatrix} x_1^{(3)} \\ x_2^{(3)} \end{pmatrix} = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} - \alpha_2 \begin{pmatrix} \dfrac{\partial f\left(x^{(2)}\right)}{\partial x_1} \\ \dfrac{\partial f\left(x^{(2)}\right)}{\partial x_2} \end{pmatrix} = \begin{pmatrix} -0{,}2853 \\ -0{,}1445 \end{pmatrix} - \\ - 0{,}24 \begin{pmatrix} 0{,}08 \\ 0{,}07264 \end{pmatrix} = \begin{pmatrix} -0{,}3045 \\ -0{,}1619 \end{pmatrix};$$

$$\left|\operatorname{grad} f\left(x^{(3)}\right)\right| = \sqrt{\left(\frac{\partial f\left(x^{(3)}\right)}{\partial x_1}\right)^2 + \left(\frac{\partial f\left(x^{(3)}\right)}{\partial x_2}\right)^2} =$$

$$= \sqrt{(-0{,}0183)^2 + (-0{,}0203)^2} = 0{,}0273 < \varepsilon = 0{,}05.$$

Заданная точность достигнута. Следовательно,

$$x^* \approx x^{(3)} = (-0{,}305;\ -0{,}162)\,; \quad f^* \approx f\left(x^{(3)}\right) = 0{,}772.$$

Можно показать, что если $f(x)$ — квадратичная функция, $f(x) = \frac{1}{2}(Qx, x) + (r, x)$, где Q — симметрическая матрица коэффициентов при квадратичных слагаемых $\left(Q = Q^T\right)$, $x = (x_1\, x_2\, \ldots\, x_n)^T$, $r = (r_1, r_2, \ldots, r_n)^T$ — коэффициенты при линейных слагаемых, (x, y) — скалярное произведение, то величина параметрического шага α_k может быть найдена в явном виде:

$$\alpha_k = \frac{\left(\operatorname{grad} f\left(x^{(k)}\right), \operatorname{grad} f\left(x^{(k)}\right)\right)}{\left(Q \operatorname{grad} f\left(x^{(k)}\right), \operatorname{grad} f\left(x^{(k)}\right)\right)}, \tag{5.7}$$

причем $\operatorname{grad} f\left(x^{(k)}\right) = Qx^{(k)} + r$.

5.4.3. Метод сопряженных направлений. Метод сопряженных направлений состоит в построении последовательности

$$x^{(k+1)} = x^{(k)} - \alpha_k p^{(k)}, \qquad k = 0, 1, 2, \ldots, \tag{5.8}$$

где α_k выбирается так же, как и в методе наискорейшего спуска:

$$\psi_k(\alpha_k) = \min_{\alpha > 0} \psi_k(\alpha), \qquad \psi_k(\alpha) = f(x^{(k)} - \alpha\, p^{(k)}),$$

а направление спуска $p^{(k)}$ — с помощью выражения

$$p^{(k)} = \operatorname{grad} f(x^{(k)}) + \beta_k p^{(k-1)}, \qquad k = 1, 2, \ldots; \tag{5.9}$$

$$\beta_0 = 0, \qquad p^{(0)} = \operatorname{grad} f(x^{(0)}); \tag{5.10}$$

$$\beta_k = \frac{\left|\operatorname{grad} f(x^{(k)})\right|^2}{\left|\operatorname{grad} f(x^{(k-1)})\right|^2} = \frac{\sum\limits_{i=1}^{n}\left[\dfrac{\partial f(x^{(k)})}{\partial x_i}\right]^2}{\sum\limits_{i=1}^{n}\left[\dfrac{\partial f(x^{(k-1)})}{\partial x_i}\right]^2}. \qquad (5.11)$$

Критерием останова является выражение (5.3) или (5.4), т. е.

$$\left|\operatorname{grad} f(x^{(k)})\right| = \left\{\sum_{i=1}^{n}\left[\frac{\partial f(x^{(k)})}{\partial x_i}\right]^2\right\}^{1/2} \leqslant \varepsilon.$$

Таким образом, метод сопряженных направлений отличается от метода наискорейшего спуска только выбором направления уменьшения функции на каждом шаге ($-p^{(k)}$ вместо $-\operatorname{grad} f(x^{(k)})$).

Отметим, что $p^{(k)}$ в (5.9) определяется не только антиградиентом $-\operatorname{grad} f(x^{(k)})$, но и направлением спуска $-p^{(k-1)}$ на предыдущем шаге. Это позволяет более полно, чем в методах градиентного и наискорейшего спуска, учитывать особенности функции $f(x)$ при построении итерационного процесса (5.8). Для квадратичных функций в E^n требуется не больше n итераций метода сопряженных направлений.

Пример 5.9. Методом сопряженных направлений найти точки минимума x^* функции $f(x) = x_1^2 + 2x_2^2 + x_1 \cdot x_2 - 7x_1 - 7x_2$.

Решение. Поскольку $f(x)$ — квадратичная функция, заданная в E^2, то точка минимума x^* может быть найдена после двух шагов метода сопряженных направлений.

Шаг 1: $k = 0$. Выбрав начальное приближение $x^{(0)} = (0;\, 0)$, по формулам (5.9)–(5.11) находим

$$p^{(0)} = \operatorname{grad} f(x^{(0)}) = (2x_1 + x_2 - 7;\ 4x_2 + x_1 - 7)\big|_{x^{(0)}} = \\ = (-7; -7)\,; \quad \psi_0(\alpha) = 98\left(2\alpha^2 - \alpha\right).$$

Из условия $\psi_0'(\alpha_0) = 0$ минимума $\psi_0(\alpha)$ получим $\alpha_0 = 0{,}25$. Отсюда

$$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix} - \alpha_0 \begin{pmatrix} p_1^{(0)} \\ p_2^{(0)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 0{,}25 \begin{pmatrix} -7 \\ -7 \end{pmatrix} = \begin{pmatrix} 7/4 \\ 7/4 \end{pmatrix}.$$

Шаг 2: $k = 1$;

$$\operatorname{grad} f\left(x^{(1)}\right) = \left(-\frac{7}{4};\ \frac{7}{4}\right),$$

откуда с учетом (5.11) имеем

$$\beta_1 = \frac{1}{16},$$

$$\begin{pmatrix} p_1^{(1)} \\ p_2^{(1)} \end{pmatrix} = \begin{pmatrix} -7/4 \\ 7/4 \end{pmatrix} + \frac{1}{16}\begin{pmatrix} -7 \\ -7 \end{pmatrix} = \begin{pmatrix} -35/16 \\ 21/16 \end{pmatrix}.$$

Поэтому $\psi_1(\alpha) = \frac{49}{32}\left(\frac{7}{2}\alpha^2 - 4\alpha - 392\right)$ и $\alpha_1 = 4/7$.

Окончательно:

$$\begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} - \alpha_1 \begin{pmatrix} p_1^{(1)} \\ p_2^{(1)} \end{pmatrix} = \begin{pmatrix} 7/4 \\ 7/4 \end{pmatrix} - \frac{4}{7}\begin{pmatrix} -35/16 \\ 21/16 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

Итак, $x^* = (x_1^*,\ x_2^*) = (3;\ 1)$; $f^* = f(x^*) = -14$.

УПРАЖНЕНИЯ.

5.1. Показать, что следующие функции на отрезке $[a, b]$ являются унимодальными:

$$\text{а) } f(x) = x^2 - 3x + x \cdot \ln x,\ \ x \in [1;\ 2];$$

$$\text{б) } f(x) = \ln\left(1 + x^2\right) - \sin x,\ \ x \in [0;\ \pi/4];$$

$$\text{в) } f(x) = \frac{1}{2}x^2 - \sin x,\ \ x \in [0;\ 1].$$

5.2. Методом перебора с точностью $\varepsilon = 0{,}05$ найти точку минимума x^* функции $f(x)$ на отрезке $[a,\ b]$ и минимум f^*:

$$\text{а) } f(x) = x^2 - 2x + e^{-x},\ x \in [1;\ 1{,}5];$$

$$\text{б) } f(x) = \operatorname{tg} x - 2\sin x,\ x \in [0;\ \pi/4];$$

$$\text{в) } f(x) = x^3 - 3\sin x,\ x \in [0{,}5;\ 1{,}0];$$

$$\text{г) } f(x) = \frac{1}{3}x^3 - 5x + x \cdot \ln x,\ x \in [1{,}5;\ 2];$$

$$\text{д) } f(x) = \sqrt{1 + x^2} + e^{-2x},\ x \in [0;\ 1].$$

5.3. Методом деления отрезка пополам с точностью ε найти точку минимума x^* функции $f(x)$ на отрезке $x \in [a,\ b]$ и минимум f^*:

$$\text{а) } f(x) = x \cdot \sin x + 2\cos x,\ x \in [-5;\ -4],\ \varepsilon = 0{,}02;$$

$$\text{б) } f(x) = 10x \cdot \ln x - \frac{x^2}{3},\ x \in [0{,}5;\ 1],\ \varepsilon = 0{,}05;$$

$$\text{в) } f(x) = 3x^4 - 10x^3 + 21x^2 + 12x,\ x \in [0;\ 0{,}5],\ \varepsilon = 0{,}01;$$

$$\text{г) } f(x) = e^x - \frac{1}{3}x^3 + 2x,\ x \in [-1{,}5;\ -1],\ \varepsilon = 0{,}01;$$

$$\text{д) } f(x) = x^6 + 3x^2 + 6x - 1,\ x \in [-1;\ 0],\ \varepsilon = 0{,}1.$$

5.4. Методом золотого сечения с точностью ε найти точку минимума x^* функции $f(x)$ на отрезке $x \in [a,\ b]$ и минимум f^*:

$$\text{а) } f(x) = x^2 + 3x(\ln x - 1),\ x \in [0{,}5;\ 1],\ \varepsilon = 0{,}05;$$

$$\text{б) } f(x) = 10x \cdot \ln x - \frac{x^2}{3},\ x \in [0{,}5;\ 1],\ \varepsilon = 0{,}05;$$

$$\text{в) } f(x) = (x + 1)^4 - 2x^2,\ x \in [-3;\ -2],\ \varepsilon = 0{,}025;$$

$$\text{г) } f(x) = x^4 + 2x^2 + 4x + 1,\ x \in [-1;\ 0],\ \varepsilon = 0{,}1;$$

д) $f(x) = x^5 - 5x^3 + 10x^2 - 5x,\ x \in [-3;\ -2],\ \varepsilon = 0{,}1.$

5.5. Методом Ньютона на всей числовой оси минимизировать функции $f(x)$ с точностью $\varepsilon = 10^{-4}$:

$$\text{а) } f(x) = 2x^2 + x + \cos^2 x;$$

$$\text{б) } f(x) = e^x + e^{-2x} + 2x;$$

$$\text{в) } f(x) = x^2 + x + \sin x;$$

$$\text{г) } f(x) = x^2 - x + e^{-x};$$

$$\text{д) } f(x) = x^2 + e^{-x}.$$

5.6. Убедиться в выпуклости следующих функций $f(x)$ на всем пространстве E^2:

$$\text{а) } f(x_1,\ x_2) = \sqrt{1 + x_1^2 + x_2^2};$$

$$\text{б) } f(x_1,\ x_2) = x_1^2 + x_2^2 - \cos\frac{x_1 - x_2}{2};$$

$$\text{в) } f(x_1,\ x_2,\ x_3) = \exp\left(x_1^2 + x_2^2 + x_3^2\right);$$

$$\text{г) } f(x_1,\ x_2,\ x_3) = 5x_1^2 + 5x_2^2 + 4x_3^2 + 4x_1x_2 - 2x_2x_3.$$

5.7. Методом градиентного спуска с точностью ε найти точку минимума x^* и минимум f^* на отрезке $x \in [a,\ b]$:

а) $f(x) = 2x_1^2 + x_2^2 + x_1x_2 + x_1 + x_2,\ x^{(0)} = (0;\ 0)$

для $\alpha_0 = 0{,}265;\ \alpha_0 = 0{,}5;$

б) $f(x) = x_1^4 + x_2^2 + x_3^2 + x_1x_3 + x_2x_3,\ x^{(0)} = (0;1;0)$

для $\alpha_0 = 0,1;\ \alpha_0 = 0,638;\ \alpha_0 = 10.$

в) $f(x) = \exp\left(x_1^2\right) + (x_1 + x_2 + x_3)^2,\ x^{(0)} = (1;\ 1;\ 1);$

$\alpha_0 = 0{,}1;\ \alpha_0 = 0{,}2127.$

5.8. Методом наискорейшего спуска с точностью $\varepsilon = 0{,}05$ найти точку минимума x^* и минимум f^*:

а) $f(x) = 2x_1^2 + x_2^2 + x_1x_2,\ x^{(0)} = (0;\ 0);$

б) $f(x) = x_1^4 + x_2^2 + x_3^2 + x_1x_3 + x_2x_3,\ x^{(0)} = (0;\ 1;\ 0);$

в) $f(x) = \exp\left(x_1^2\right) + (x_1 + x_2 + x_3)^2,\ x^{(0)} = (0;\ 1;\ 0);$

г) $f(x) = x_1^2 + 4x_1x_2 + 17x_2^2 + 5x_2,\ x^{(0)} = (0;\ 1;\ 0).$

5.9. Методом сопряженных направлений с точностью $\varepsilon = 0{,}001$ найти точку минимума x^* и минимум f^*:

а) $f(x) = x_1 + 5x_2 + \exp\left(x_1^2 + x_2^2\right);$

б) $f(x) = x_1^2 + 3x_2^2 + \cos(x_1 + x_2);$

в) $f(x) = x_1^2 + \exp\left(x_1^2 + x_2^2\right) + 4x_1 + 3x_2;$

г) $f(x) = x_1^4 + 2x_2^4 + x_1^2x_2^2 + 2x_1 + x_2;$

д) $f(x) = \exp\left(x_1^2 + x_2^2\right) + \ln\left(4 + x_2^2 + 2x_3^2\right);$

е) $f(x) = x_1 + x_2 - 5x_3 + \exp\left(x_1^2 + 2x_2^2 + x_3^2\right).$

Часть II

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ ДЛЯ УРАВНЕНИЙ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Наиболее распространенными численными методами решения задач математической физики в настоящее время являются следующие: метод конечных разностей (МКР), метод конечных элементов (МКЭ), метод граничных элементов (МГЭ). Ниже наиболее подробно рассматриваются методы конечных разностей и конечных элементов.

ГЛАВА VI

МЕТОД КОНЕЧНЫХ РАЗНОСТЕЙ

Программа

Постановка задач математической физики для уравнений различных типов. Основные определения при замене дифференциальных операторов конечно-разностными: сетка, сеточная функция, шаблон, временной слой, явная и неявная конечно-разностные схемы. Конечно-разностные схемы для уравнений математической физики различных типов. Основные понятия, связанные с конечно-разностной аппроксимацией дифференциальных задач: аппроксимация и порядок аппроксимации, устойчивость, сходимость и порядок сходимости, консервативность и корректность. Теорема эквивалентности. Анализ порядка аппроксимации. Исследование устойчивости. Метод гармонического анализа, принцип максимума, спектральный метод, энергетический метод для явных и неявных схем. Метод конечных разностей решения параболических задач, однородные и консервативные схемы. Неявно-явные схемы, схема Кранка–Николсона. Методы прямых. Ме-

тод конечных разностей решения эллиптических задач, явные, неявные, неявно-явные схемы. Конечно-разностная аппроксимация краевых условий, содержащих производные. Метод установления и его обоснование для эллиптических задач. Разностно-итерационный метод Либмана.

§ 6.1. Постановка задач математической физики

Чтобы поставить задачу математической физики, необходимо вывести дифференциальное уравнение в частных производных, описывающее рассматриваемый физический процесс, а также начальные и краевые условия. При этом начальные условия ставятся для уравнений, содержащих частные производные по времени (уравнения описывают нестационарные физические процессы). Краевые (граничные) условия ставятся для уравнений, описывающих физические процессы в ограниченных областях.

Задачи математической физики, содержащие начальные и краевые условия, называются *начально-краевыми*; задачи, содержащие только граничные условия — *краевыми*, а задачи, содержащие только начальные условия (в бесконечных областях) — *задачами Коши*.

Известно, что количество условий и вид начальных и краевых условий зависят от типов уравнений математической физики, среди которых различают параболический, гиперболический и эллиптический типы, количества границ, количества разрывов в граничных условиях, порядка дифференциальных уравнений.

Вывод основных уравнений математической физики, начальных и краевых условий к ним дается в курсе «Уравнения математической физики». Здесь же ограничимся математической формулировкой типовых задач математической физики.

6.1.1. Постановка задач для уравнений параболического типа. Классическим примером уравнения параболического типа является уравнение теплопроводности (диффузии). В одномерном по пространству случае однородное (без источников энергии) уравнение теплопроводности имеет вид

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0. \tag{6.1}$$

Если на границах x =0 и x = l заданы значения искомой функции $u(x,t)$ *в виде*

$$u(0,t) = \varphi_1(t), \quad x = 0, \quad t > 0, \tag{6.2}$$

$$u(l,t) = \varphi_2(t), \quad x = l, \quad t > 0, \tag{6.3}$$

т. е. *граничные условия первого рода*, и, кроме того, заданы начальные условия

$$u(x,0) = \psi(x), \quad 0 \leqslant x \leqslant l, \quad t = 0, \tag{6.4}$$

то задачу (6.1)–(6.4) называют *первой начально-краевой задачей для уравнения теплопроводности* (6.1).

В терминах теории теплообмена $u(x,t)$ — распределение температуры в пространственно-временной области $\Omega \times \mathrm{T} = \{0 \leqslant \mathrm{x} \leqslant l;\ 0 \leqslant t \leqslant T\}$, a^2 — коэффициент температуропроводности, а (6.2), (6.3) с помощью функций $\varphi_1(t)$, $\varphi_2(t)$ задают температуру на границах $x = 0$ и $x = l$.

Если на границах x = 0 и x = l заданы значения производных искомой функции по пространственной переменной:

$$\frac{\partial u(0,t)}{\partial x} = \varphi_1(t), \quad x = 0, \quad t > 0, \tag{6.5}$$

$$\frac{\partial u(l,t)}{\partial x} = \varphi_2(t), \quad x = l, \quad t > 0, \tag{6.6}$$

т. е. *граничные условия второго рода*, то задачу (6.1), (6.5), (6.6), (6.4) называют второй начально-краевой задачей для уравнения теплопроводности (6.1). В терминах теории теплообмена на границах в этом случае заданы тепловые потоки.

Если на границах заданы линейные комбинации искомой функции и ее производной по пространственной переменной:

$$\frac{\partial u(0,t)}{\partial x} + \alpha\, u(0,t) = \varphi_1(t), \quad x = 0, \quad t > 0, \tag{6.7}$$

$$\frac{\partial u(l,t)}{\partial x} + \beta\, u(l,t) = \varphi_2(t), \quad x = l, \quad t > 0, \tag{6.8}$$

т.е. *граничные условия третьего рода*, то задачу (6.1), (6.7), (6.8), (6.4) называют третьей начально-краевой задачей для уравнения теплопроводности (6.1). В терминах теории теплообмена granич-

ные условия (6.7), (6.8) задают теплообмен между газообразной или жидкой средой с известными температурами $\varphi_1(t)/\alpha$ на границе x = 0 и $\varphi_2(t)/\beta$ на границе $x = l$ и границами расчетной области с неизвестными температурами $u(0,t)$, $u(l,t)$. Коэффициенты α, β — известные коэффициенты теплообмена между газообразной или жидкой средой и соответствующей границей.

Для пространственных задач теплопроводности в области $\overline{\Omega} = \Omega + \Gamma$ первая начально-краевая задача имеет вид

$$\begin{cases} \dfrac{\partial u}{\partial t} = a^2\left(\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial^2 u}{\partial z^2}\right), \ M(x,y,z) \in \Omega, \ t > 0; & (6.9) \\ u(M,t) = \varphi(M,t), \quad M(x,y,z) \in \Gamma, \quad t > 0; & (6.10) \\ u(M,0) = \psi(M), \quad M(x,y,z) \in \overline{\Omega}, \quad t > 0. & (6.11) \end{cases}$$

Аналогично ставится вторая и третья начально-краевые задачи для пространственного уравнения (6.9).

На практике часто ставятся начально-краевые задачи теплопроводности со смешанными краевыми условиями, когда на границах задаются граничные условия различных родов.

6.1.2. Постановка задач для уравнений гиперболического типа. Классическим примером уравнения гиперболического типа является волновое уравнение, которое в области $0 < < x < l$, $t > 0$ имеет вид

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t = 0.$$

Здесь $u(x,t)$ — малые продольные или поперечные перемещения (колебания) стержня, a — скорость звука в материале, из которого изготовлен стержень.

Если концы стержня движутся по заданным законам, то есть на концах заданы перемещения (или значения искомой функции), *то первая начально-краевая задача для волнового уравнения имеет вид* (причем, если концы стержня жестко закреплены, то $\varphi_1(t) = \varphi_2(t) = 0$)

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = a^2 \dfrac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0; & (6.12)\\ u(0,t) = \varphi_1(t), \quad x = 0, \quad t > 0; & (6.13)\\ u(l,t) = \varphi_2(t), \quad x = l, \quad t > 0; & (6.14)\\ u(x,0) = \psi_1(x), \quad 0 \leqslant x \leqslant l, \quad t = 0; & (6.15)\\ \dfrac{\partial u(x,0)}{\partial t} = \psi_2(x), \quad 0 \leqslant x \leqslant l, \quad t = 0. & (6.16) \end{cases}$$

Как видно, в задачах для волнового уравнения, кроме начального распределения искомой функции, задается еще распределение начальной скорости перемещения.

Если на концах стержня заданы значения силы, которая по закону Гука пропорциональна значениям производной перемещения по пространственной переменной (т. е. на концах заданы значения первых производных по переменной x), то ставится *вторая начально-краевая задача для волнового уравнения*:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = a^2 \dfrac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0;\\ \dfrac{\partial u(0,t)}{\partial x} = \varphi_1(t), \quad x = 0, \quad t > 0;\\ \dfrac{\partial u(l,t)}{\partial x} = \varphi_2(t), \quad x = l, \quad t > 0;\\ u(x,0) = \psi_1(x), \quad 0 \leqslant x \leqslant l, \quad t = 0;\\ \dfrac{\partial u(x,0)}{\partial t} = \psi_2(x), \quad 0 \leqslant x \leqslant l, \quad t = 0. \end{cases}$$

В условиях, когда концы стержня *свободны*, функции $\varphi_1(t) = \varphi_2(t) = 0$.

Наконец, в условиях, когда концы закреплены *упруго*, т. е. на концевые заделки действуют силы, пропорциональные перемещениям, ставится *третья начально-краевая задача для волнового уравнения*:

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0;$$

$$\frac{\partial u(0,t)}{\partial x} + \alpha u(0,t) = \varphi_1(t), \quad x = 0, \quad t > 0;$$

$$\frac{\partial\, u(l,t)}{\partial\, x} + \beta u(l,t) = \varphi_2(t), \qquad x = l, \qquad t > 0;$$

$$u(x,0) = \psi_1(x), \qquad 0 \leqslant x \leqslant l, \qquad t = 0;$$

$$\frac{\partial\, u(x,0)}{\partial\, t} = \psi_2(x), \qquad 0 \leqslant x \leqslant l, \qquad t = 0.$$

Аналогично ставятся двумерные и трехмерные начально-краевые задачи для двумерного и трехмерного волнового уравнения.

Необходимо отметить, что волновое уравнение легко трансформируется в систему уравнений *акустики*, являющихся простейшей линейной моделью газодинамических течений.

Действительно, введем следующие обозначения:

$$\begin{cases} \dfrac{\partial\, u}{\partial\, t} = p, & (6.17) \\[2ex] \dfrac{\partial\, u}{\partial\, x} = v. & (6.18) \end{cases}$$

Продифференцируем (6.17) по переменной t, а (6.18) — по переменной x, результат подставим в волновое уравнение, получим

$$\frac{\partial\, p}{\partial\, t} = a^2 \frac{\partial\, v}{\partial\, x}. \tag{6.19}$$

Продифференцируем затем (6.17) по x, а (6.18) — по t и, полагая, что смешанные производные от $u(x,t)$ по переменным x и t не зависят от порядка дифференцирования, получим

$$\frac{\partial\, v}{\partial\, t} = \frac{\partial\, p}{\partial\, x}. \tag{6.20}$$

Здесь p, v — возмущения (слабые колебания) давления и скорости в акустической волне, а систему (6.19), (6.20) называют уравнениями акустики. Поскольку (6.19), (6.20) — линейное приближение уравнений газовой динамки, они используются для исследования устойчивости численных методов решения уравнений газовой динамики.

6.1.3. Постановка задач для уравнений эллиптического типа. Классическим примером уравнения эллиптического типа является уравнение Пуассона

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$$

или уравнение Лапласа при $f(x, y) \equiv 0$.

Здесь функция $u(x, y)$ может иметь различный физический смысл, например: стационарное, независящее от времени, распределение температуры, скорость потенциального (безвихревого) течения идеальной (без трения и теплопроводности) жидкости, распределение напряженностей электрического и магнитного полей, потенциала в силовом поле тяготения и т. п.

Если на границе Γ расчетной области $\overline{\Omega} = \Omega + \Gamma$ задана искомая функция, то соответствующая *первая краевая* задача для уравнения Лапласа или Пуассона называется *задачей Дирихле*:

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = f(x, y), \qquad (x, y) \in \Omega; & (6.21) \\[2ex] u(x, y)|_{\Gamma} = \varphi(x, y), \qquad (x, y) \in \Gamma. & (6.22) \end{cases}$$

Если на границе Γ задается нормальная производная искомой функции, то соответствующая *вторая краевая* задача называется задачей Неймана для уравнения Лапласа или Пуассона:

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = f(x, y), \qquad (x, y) \in \Omega; & (6.23) \\[2ex] \left.\dfrac{\partial\, u(x, y)}{\partial\, n}\right|_{\Gamma} = \varphi(x, y), \qquad (x, y) \in \Gamma. & (6.24) \end{cases}$$

При этом n — направление внешней к границе Γ нормали.

Более приемлемой является координатная форма краевого условия (6.24):

$$\frac{\partial u}{\partial x} \cos(\widehat{n, i}) + \frac{\partial u}{\partial y} \cos(\widehat{n, j}) = \varphi(x, y),$$

где $\cos(\widehat{n, i})$, $\cos(\widehat{n, j})$ — направляющие косинусы внешнего вектора единичной нормали к границе Γ, i и j — орты базисных векторов.

Наконец третья краевая задача для уравнения Пуассона (Лапласа) имеет вид

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = f(x, y), \qquad (x, y) \in \Omega; \\ \left.\dfrac{\partial u(x, y)}{\partial n}\right|_\Gamma + \alpha\, u|_\Gamma = \varphi(x, y), \qquad (x, y) \in \Gamma. \end{cases}$$

Замечание: следует отметить, что в вышеперечисленных постановках *число начальных условий равно порядку дифференциального уравнения по времени*, а старший порядок производной по времени в начальных условиях на единицу меньше порядка дифференциального уравнения по времени.

Старший порядок производной по пространственной переменной в краевых условиях равен порядку дифференциального уравнения по пространственной переменной минус единица. Количество краевых условий для многомерных задач не ограничено, так как на разных участках границы могут быть заданы граничные условия различных родов.

В одномерных задачах с одной пространственной переменной количество граничных условий точно равно порядку дифференциального уравнения по пространственной переменной.

§ 6.2. Основные определения и конечно-разностные схемы для различных задач математической физики

6.2.1. Основные определения. Основные определения, связанные с методом конечных разностей, рассмотрим на примере конечно-разностного решения первой начально-краевой задачи для уравнения теплопроводности (6.1)–(6.4). В этом же параграфе будут рассмотрены простейшие конечно-разностные схемы для задач математической физики, содержащих дифференциальные уравнения различных типов.

Нанесем на пространственно-временную область $0 \leqslant x \leqslant l$, $0 \leqslant t \leqslant T$ конечно-разностную сетку $\omega_{h,\tau}$:

$$\omega_{h,\tau} = \{x_j = jh,\ j = \overline{0, N};\ t^k = k\tau,\ k = \overline{0, K}\}, \tag{6.25}$$

с пространственным шагом $h = l/N$ и шагом по времени $\tau = T/K$ (рис. 6.1).

Введем два *временных слоя*: нижний $t^k = k\tau$, на котором распределение искомой функции $u(x_j, t^k)$, $j = \overline{0, N}$, известно (при

$k = 0$ распределение определяется начальным условием (6.4) $u(x_j, t^0) = \psi(x_j)$), и верхний временной слой $t^{k+1} = (k+1)\tau$, на котором распределение искомой функции $u(x_j, t^{k+1})$, $j = \overline{0, N}$, подлежит определению.

Сеточной функцией задачи (6.1)–(6.4) (обозначение u_j^k) назовем однозначное отображение *целых* аргументов j, k в значения функции $u_j^k = u(x_j, t^k)$.

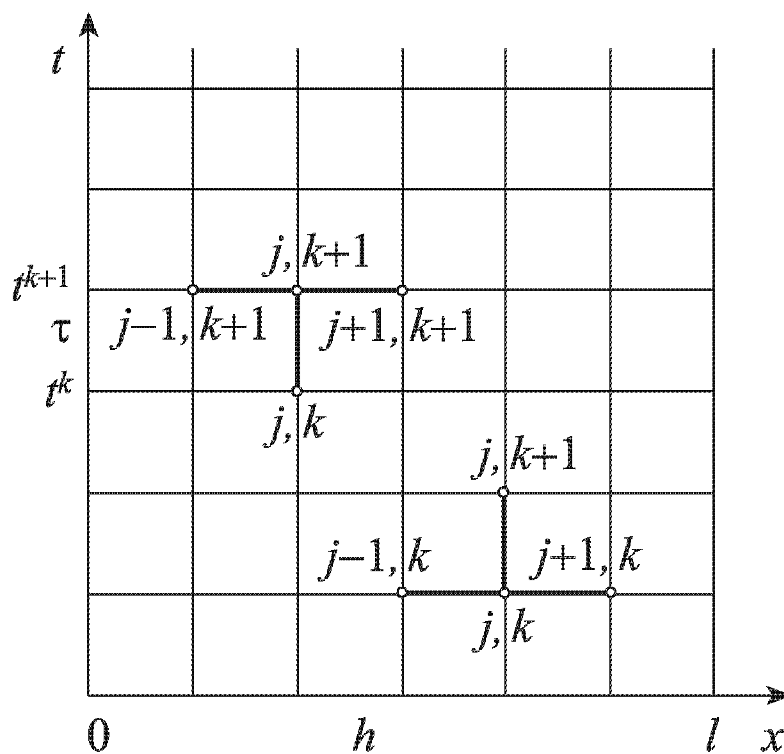


Рис. 6.1. Конечно-разностная сетка

На введенной сетке (6.25) введем сеточные функции u_j^k, u_j^{k+1}, первая из которых известна, вторая — подлежит определению. Для ее определения в задаче (6.1)–(6.4) заменим (аппроксимируем) дифференциальные операторы отношением конечных разностей (см. § 3.4 «Численное дифференцирование»), получим

$$\left.\frac{\partial u}{\partial t}\right|_j^k = \frac{u_j^{k+1} - u_j^k}{\tau} + O(\tau), \tag{6.26}$$

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_j^k = \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{h^2} + O(h^2). \tag{6.27}$$

Подставляя (6.26), (6.27) в задачу (6.1)–(6.4), получим *явную конечно-разностную схему* для этой задачи в форме

$$\begin{gathered} \frac{u_j^{k+1} - u_j^k}{\tau} = a^2 \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{h^2} + O(\tau + h^2), \\ j = \overline{1, N-1}, \quad k = \overline{0, K-1}; \\ u_0^{k+1} = \varphi_1\left(t^{k+1}\right), \quad u_N^{k+1} = \varphi_2\left(t^{k+1}\right), \quad k = \overline{0,\ K}; \\ u_j^0 = \psi(x_j), \quad j = \overline{0, N}. \end{gathered} \tag{6.28}$$

В каждом уравнении этой задачи все значения сеточной функции известны, за исключением одного, u_j^{k+1}, которое может быть

определено *явно* из соотношений (6.28). В соотношения (6.28) краевые условия входят при значениях $j = 1$ и $j = N - 1$, а начальное условие — при $k = 0$.

Если в (6.27) дифференциальный оператор по пространственной переменной аппроксимировать отношением конечных разностей на верхнем временном слое:

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_j^{k+1} = \frac{u_{j+1}^{k+1} - 2u_j^{k+1} + u_{j-1}^{k+1}}{h^2} + O(h^2), \tag{6.29}$$

то после подстановки (6.26), (6.29) в задачу (6.1)–(6.4) получим *неявную конечно-разностную схему* для этой задачи:

$$\begin{gathered} \frac{u_j^{k+1} - u_j^k}{\tau} = a^2 \frac{u_{j+1}^{k+1} - 2u_j^{k+1} + u_{j-1}^{k+1}}{h^2} + O(\tau + h^2), \\ j = \overline{1, N-1}, \quad k = \overline{0, K-1}; \\ u_0^{k+1} = \varphi_1(t^{k+1}), \quad u_N^{k+1} = \varphi_2(t^{k+1}), \quad k = \overline{0,\ K-1}; \\ u_j^0 = \psi(x_j), \quad j = \overline{0, N}. \end{gathered} \tag{6.30}$$

Теперь сеточную функцию u_j^{k+1} на верхнем временном слое можно получить из решения СЛАУ (6.30) с трехдиагональной матрицей. Эта СЛАУ в форме, пригодной для использования метода прогонки, имеет вид

$$\begin{array}{l} a_1 = 0; \\ \\ c_{N-1} = 0; \end{array} \left\{ \begin{array}{ll} b_1 u_1^{k+1} + c_1 u_2^{k+1} = d_1, & j = 1, \\ a_j u_{j-1}^{k+1} + b_j u_j^{k+1} + c_j u_{j+1}^{k+1} = d_j, & j = \overline{2, N-2}, \\ a_{N-1} u_{N-2}^{k+1} + b_{N-1} u_{N-1}^{k+1} = d_{N-1}, & j = N-1, \end{array} \right.$$

где $a_j = \sigma$, $j = \overline{2, N-1}$; $b_j = -(1 + 2\sigma)$, $j = \overline{1, N-1}$; $c_j = \sigma$, $j = \overline{1, N-2}$; $d_j = -u_j^k$, $j = \overline{2, N-2}$; $d_1 = -(u_1^k + \sigma\varphi_1(t^{k+1}))$; $d_{N-1} = -(u_{N-1}^k + \sigma\varphi_2(t^{k+1}))$; $\sigma = \dfrac{a^2\tau}{h^2}$.

Шаблоном конечно-разностной схемы называют ее геометрическую интерпретацию на конечно-разностной сетке.

На рис. 6.2 приведены шаблоны для явной (6.28) и неявной (6.30) конечно-разностных схем при аппроксимации задачи (6.1)–(6.4).

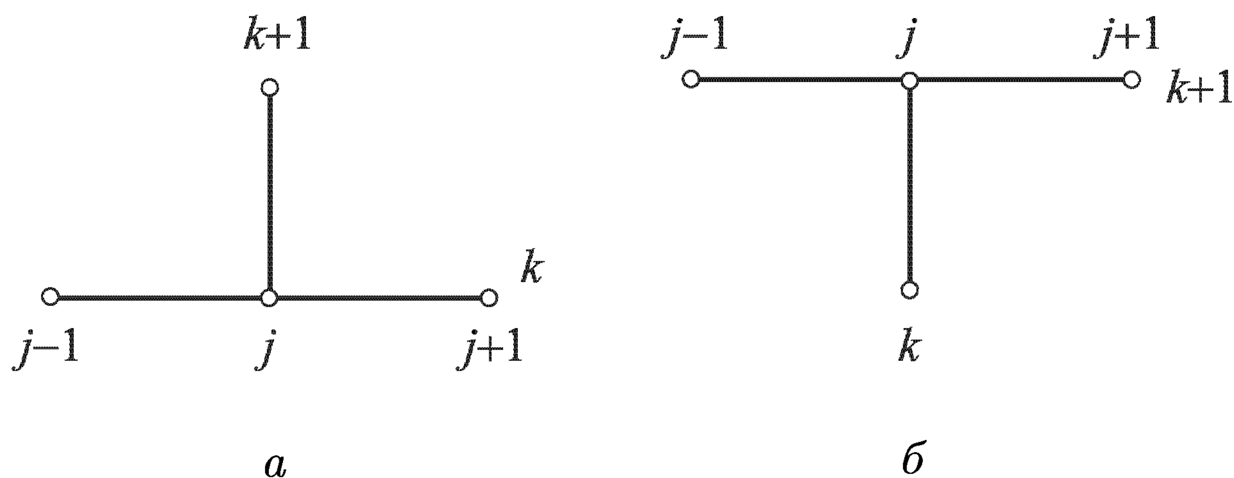


Рис. 6.2. Шаблоны явной (*а*) и неявной (*б*) конечно-разностных схем для уравнения теплопроводности

6.2.2. Конечно-разностная аппроксимация задач для уравнений гиперболического типа. Рассмотрим первую начально-краевую задачу для волнового уравнения (6.12)–(6.16). На пространственно-временной сетке (6.25) будем аппроксимировать дифференциальное уравнение (6.12) одной из следующих конечно-разностных схем:

$$\frac{u_j^{k+1} - 2u_j^k + u_j^{k-1}}{\tau^2} = a^2 \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{h^2} + O\left(\tau^2 + h^2\right),$$

$$j = \overline{1, N-1}; \quad k = 1, 2, \ldots \qquad (6.31)$$

с шаблоном на рис. 6.3 *а* и

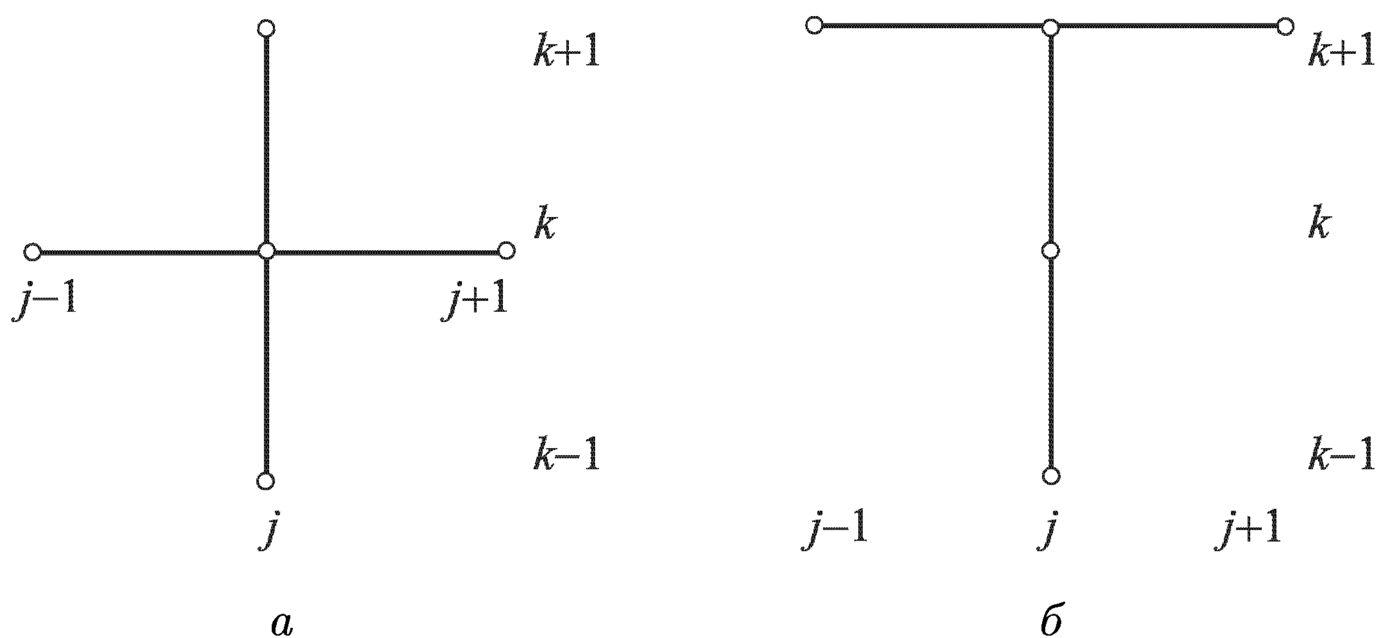


Рис. 6.3. Шаблоны конечно-разностных схем для волнового уравнения

$$\frac{u_j^{k+1} - 2u_j^k + u_j^{k-1}}{\tau^2} = a^2 \frac{u_{j+1}^{k+1} - 2u_j^{k+1} + u_{j-1}^{k+1}}{h^2} + O\left(\tau + h^2\right),$$

$$j = \overline{1, N-1}; \quad k = 1, 2, \ldots \qquad (6.32)$$

с шаблоном на рис. 6.3 *б*.

При этом схема (6.31) является *явной*. С ее помощью решение u_j^{k+1}, $j = \overline{1, N-1}$, $k = 1, 2, \ldots$, определяется сразу, поскольку значения сеточных функций u_j^{k-1}, u_j^k на нижних временных слоях должны быть известны. В соответствии с шаблоном для этой схемы порядок аппроксимации равен двум как по пространственной, так и по временной переменной (достаточно разложить на точном решении нецентральные значения сеточной функции u_j^{k-1},u_j^{k+1}, u_{j-1}^k, u_{j+1}^k в ряд Тейлора в окрестности центрального узла (x_j, t^k) до производных четвертого порядка включительно соответственно по переменным t и x).

Схема (6.32) является *неявной схемой*, ее можно свести к СЛАУ с трехдиагональной матрицей, реализуемой методом прогонки. В обеих схемах необходимо знать значения u_j^{k-1}, u_j^k, $j = \overline{1, N-1}$, $k = 1, 2, \ldots$, на нижних временных слоях. Для $k = 1$ это делается следующим образом:

$$u_j^0 = \psi_1(x_j), \quad j = \overline{0, N}, \tag{6.33}$$

где $\psi_1(x)$ — функция из начального условия (6.15).

Для определения u_j^1 можно воспользоваться простейшей аппроксимацией второго начального условия (6.16):

$$\frac{u_j^1 - u_j^0}{\tau} = \psi_2(x_j).$$

Отсюда для искомых значений u_j^1 получаем следующее выражение:

$$u_j^1 = \psi_1(x_j) + \psi_2(x_j)\tau.$$

Недостатком такого подхода является первый порядок аппроксимации второго начального условия. Для повышения порядка аппроксимации воспользуемся следующей процедурой.

Разложим u_j^1 в ряд Тейлора на точном решении по времени в окрестности $t = 0$:

$$u_j^1 = u(x_j, 0 + \tau) = u_j^0 + \left.\frac{\partial u}{\partial t}\right|_j^0 \tau + \left.\frac{\partial^2 u}{\partial t^2}\right|_j^0 \frac{\tau^2}{2} + O(\tau^3).$$

Для определения второй производной в полученном выражении воспользуемся исходным дифференциальным уравнением

$$\left.\frac{\partial^2 u}{\partial t^2}\right|_j^0 = a^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_j^0 = a^2 \psi_1''(x_j).$$

В результате получаем искомое значение сеточной функции u_j^1 со вторым порядком точности:

$$u_j^1 = \psi_1(x_j) + \psi_2(x_j)\tau + a^2\psi_1''(x_j)\frac{\tau^2}{2}. \qquad (6.34)$$

Для определения u_j^2 из схемы (6.31) или (6.32) достаточно (6.33), (6.34) подставить в соответствующие схемы. При $k = 2$ значения u_j^1, u_j^2 сеточной функции известны, а u_j^3 определяются из схемы (6.31) или (6.32) и т. д.

Краевые условия (6.13) и (6.14) в схемах (6.31), (6.32) используются автоматически соответственно при $j = 1$ и $j = N - 1$, так же как в схемах (6.28), (6.30) для задачи теплопроводности.

6.2.3. Конечно-разностная аппроксимация задач для уравнений эллиптического типа. Рассмотрим краевую задачу для уравнений Лапласа или Пуассона (6.21), (6.22) в прямоугольнике $x \in \in [0, l_1], y \in [0, l_2]$, на который наложим сетку

$$\omega_{h_1,h_2} = \{x_i = ih_1, i = \overline{0, N_1}; \\ y_j = jh_2, j = \overline{0, N_2}\}. \qquad (6.35)$$

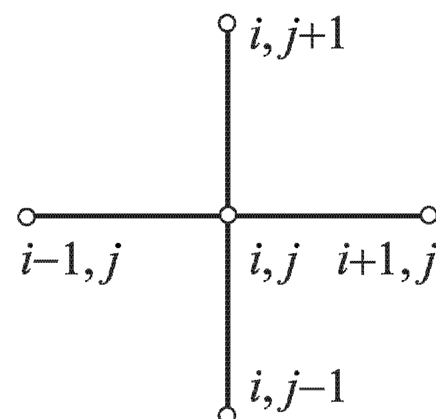


Рис. 6.4. Центрально-симметричный шаблон для уравнения Лапласа

На этой сетке аппроксимируем задачу (6.21), (6.22) с помощью отношения конечных разностей по следующей схеме (вводится сеточная функция u_{ij}, $i = \overline{0, N_1}$, $j = \overline{0, N_2}$):

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h_1^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h_2^2} + O\left(h_1^2 + h_2^2\right) =$$

$$= f(x_i, y_j), \quad i = \overline{1, N_1 - 1}, \quad j = \overline{1, N_2 - 1}, \qquad (6.36)$$

которая на шаблоне (6.4) имеет второй порядок по переменным x и y, поскольку шаблон центрально симметричен.

СЛАУ (6.36) имеет пятидиагональный вид. Решать ее можно различными методами линейной алгебры, например методом Гаусса, итерационными методами, методом матричной прогонки и т. п.

Рассмотрим разностно-итерационный метод Либмана численного решения задачи Дирихле (6.21), (6.22). Для простоты изложения этого метода примем $h_1 = h_2 = h$, тогда из схемы (6.36) получим (n — номер итерации)

$$u_{i,j}^{(n+1)} = \frac{1}{4}\left[u_{i+1,j}^{(n)} + u_{i-1,j}^{(n)} + u_{i,j-1}^{(n)} + u_{i,j+1}^{(n)} - h^2 \cdot f_{i,j}\right],$$

$$f_{i,j} = f(x_i, y_j), \quad i = \overline{1, N_1 - 1}, \;\; j = \overline{1, N_2 - 1}. \quad (6.37)$$

На каждой координатной линии (например, $y_j = \text{const}$, $j = \overline{1, N_2 - 1}$) с помощью линейной интерполяции (см. рис. 6.5)

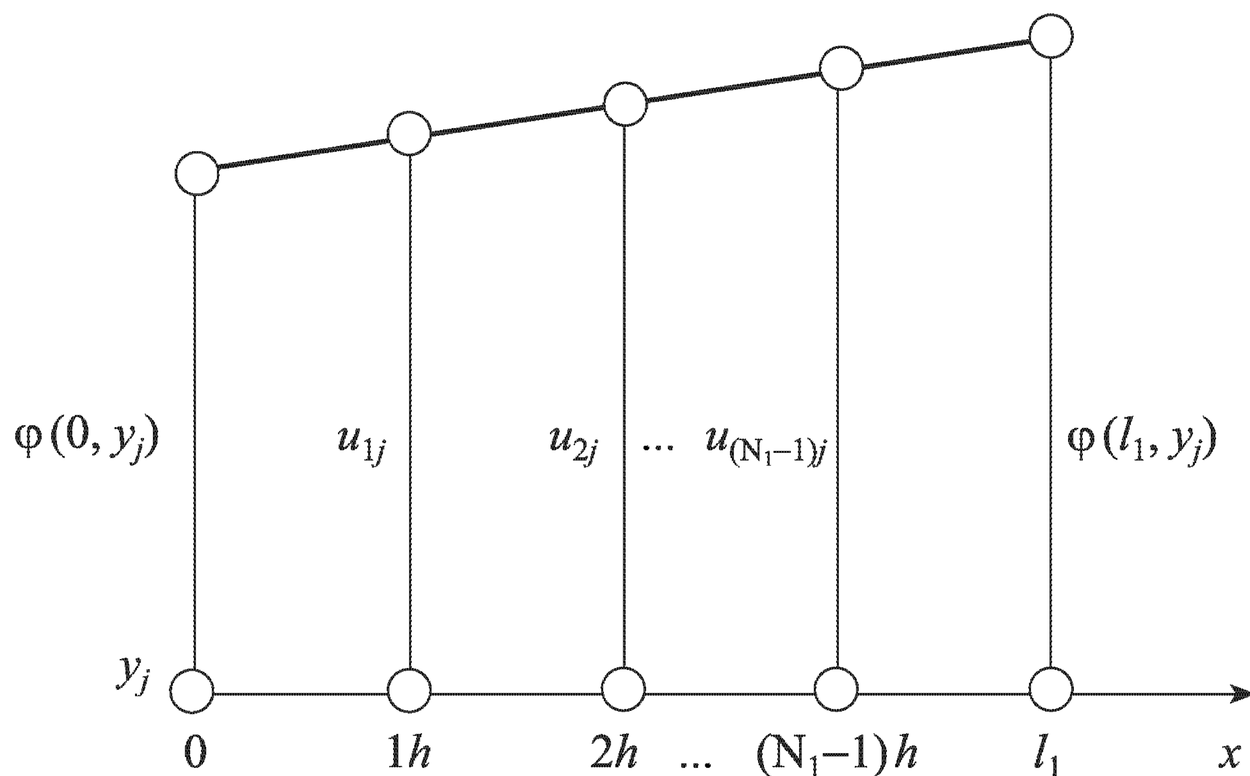


Рис. 6.5. К разностно-итерационному методу Либмана

граничных значений $\varphi(x, y)$ определим $u_{i,j}^{(0)}$ на нулевой итерации, подставив которые в (6.37), получим распределение $u_{i,j}^{(1)}$ на первой итерации:

$$u_{i,j}^{(1)} = \frac{1}{4}\left[u_{i+1,j}^{(0)} + u_{i-1,j}^{(0)} + u_{i,j+1}^{(0)} + u_{i,j-1}^{(0)} - h^2 \cdot f_{i,j}\right],$$

$$i = \overline{1, N_1 - 1}, \quad j = \overline{1, N_2 - 1}.$$

Это распределение $u_{i,j}^{(1)}$ снова подставляется в (6.37), получаем распределение $u_{i,j}^{(2)}$ и т. д. Процесс Либмана прекращается, когда

$$\left\| u^{(n+1)} - u^{(n)} \right\| \leqslant \varepsilon, \quad \left\| u^{(n)} \right\| = \max_{i,j} \left| u_{i,j}^{(n)} \right|,$$

где ε — наперед заданная точность.

Замечание. Метод простых итераций для решения СЛАУ, возникающих при аппроксимации уравнения Пуассона (Лапласа), отличается довольно медленной сходимостью. Этот недостаток может стать существенным при использовании мелких сеток, когда число уравнений в системе становится большим. С более эффективными методами решения таких СЛАУ можно познакомиться, например, в [2].

§ 6.3. Основные понятия, связанные с конечно-разностной аппроксимацией дифференциальных задач

К основным понятиям, связанным с методом конечных разностей, относятся следующие: аппроксимация, порядок аппроксимации, устойчивость, сходимость, порядок сходимости (или точность), консервативность и корректность. Определим каждое из этих понятий.

6.3.1. Аппроксимация и порядок аппроксимации. Запишем дифференциальную задачу в операторной форме

$$LU = f,$$

где L — один из дифференциальных операторов:

$$L = \begin{cases} \dfrac{\partial}{\partial t} - a^2\dfrac{\partial^2}{\partial x^2} \text{ — \textit{диффузионный};} \\[2mm] \dfrac{\partial^2}{\partial t^2} - a^2\dfrac{\partial^2}{\partial x^2} \text{ — \textit{волновой};} \\[2mm] \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2} \text{ — \textit{лапласиан};} \end{cases}$$

$U(x, y)$ — искомая функция, удовлетворяющая дифференциальной задаче; f — входные данные (т. е. начальные и краевые условия, правые части и т. п.). Операторная форма $(LU)_h = f_h$ описывает дифференциальную задачу в узлах сетки, а операторная форма $L_hU_h = f_h$ описывает конечно-разностную схему на точном решении $U(x, t)$, т. е. в конечно-разностной схеме вместо значений сеточной функции подставлены точные (неизвестные) значения искомой функции. Для перечисленных дифференциальных операторов L конечно-разностные операторы L_h имеют вид

$$L = \begin{cases} \dfrac{\Delta}{\tau} - a^2\dfrac{\Delta^2}{h^2} \text{ — \textit{диффузионный}}; \\ \dfrac{\Delta^2}{\tau^2} - a^2\dfrac{\Delta^2}{h^2} \text{ — \textit{волновой}}; \\ \dfrac{\Delta^2}{h_1^2} + \dfrac{\Delta^2}{h_2^2} \text{ — \textit{лапласиан}}. \end{cases}$$

Наконец, операторная форма конечно-разностной схемы (например, схемы (6.28) или (6.30)) имеет вид

$$L_h u_h = f_h. \tag{6.38}$$

Введем норму сеточной функции с помощью выражения

$$\left\|u^k\right\| = \max_j \left|u_j^k\right|, \quad k = 0, 1, 2, \ldots \tag{6.39}$$

Определение. *Конечно-разностная схема* (6.38) *аппроксимирует дифференциальную задачу на точном решении, если какая-либо норма разности* (*не обязательно в виде* (6.39)) $\|(LU)_h - L_h U_h\|$ *стремится к нулю при* τ, $h \to 0$:

$$\|(LU)_h - L_h U_h\| \xrightarrow[\tau, h \to 0]{} 0. \tag{6.40}$$

Определение. *Конечно-разностная схема* (6.38) *аппроксимирует дифференциальную задачу на точном решении с порядком* p *по времени и порядком* q *по пространственной переменной, если какая-либо норма разности* $\|(LU)_h - L_h U_h\|$ *удовлетворяет равенству*

$$\|(LU)_h - L_h U_h\| = O\left(\tau^p + h^q\right). \tag{6.41}$$

Таким образом, если конечно-разностная схема аппроксимирует дифференциальную задачу, то речь идет *о близости* дифференциального и конечно-разностного операторов в узлах сетки.

6.3.2. Устойчивость. Пусть в конечно-разностной схеме (6.38) входные данные f_h получили возмущения и приняли значения $\widetilde{f}_h$. Тогда сеточная функция u_h также получит возмущение и примет значение $\widetilde{u}_h$.

Определение. *Конечно-разностная схема* (6.38) *устойчива по входным данным, если найдется такая ограниченная константа* $K > 0$, *не зависящая от сеточных характеристик* τ,

h и входных данных f_h, что выполняется неравенство

$$\|u_h - \widetilde{u}_h\| \leqslant K \left\| f_h - \widetilde{f}_h \right\|. \tag{6.42}$$

Таким образом, понятие устойчивости интерпретируется следующим образом: конечно-разностная схема устойчива, если для малых возмущений входных данных (начально-краевых условий и правых частей) конечно-разносная схема (6.38) обеспечивает малые возмущения сеточной функции u_h, т. е. решение с помощью конечно-разностной схемы находится под контролем входных данных.

Если во входные данные f_h входят только начальные условия или только краевые условия, или только правые части, то говорят *об устойчивости* соответственно *по начальным условиям*, *по краевым условиям* или *по правым частям.*

Определение. *Конечно-разностная схема* (6.38) ***абсолютно*** (***безусловно***) *устойчива, если неравенство* (6.42) *выполняется при любых значениях сеточных характеристик τ и h, т. е. на шаги сетки не накладывается никаких ограничений.*

Определение. *Конечно-разностная схема* (6.38) ***условно устойчива***, *если неравенство* (6.42) *выполняется для сеточных характеристик τ и h, на которые накладываются определенные ограничения.*

6.3.3. Сходимость и порядок сходимости. ***Определение.*** *Решение u_h, полученное с помощью конечно-разностной схемы* (6.38), *сходится к точному решению U, если какая-либо норма разности $\|U - u_h\|$ стремится к нулю при стремлении к нулю сеточных характеристик τ, h:*

$$\|U - u_h\| \xrightarrow[\tau, h \to 0]{} 0. \tag{6.43}$$

Определение. *Конечно-разностная схема* (6.38) *имеет p-й порядок сходимости* (*порядок точности*) *по времени и q-й порядок сходимости по пространственной переменной, если какая-либо норма разности $\|U - u_h\|$ удовлетворяет равенству*

$$\|U - u_h\| = O\left(\tau^p + h^q\right). \tag{6.44}$$

Таким образом, порядок сходимости (порядок точности) характеризует близость конечно-разностного и точного (неизвестного) решения.

6.3.4. Теорема эквивалентности о связи аппроксимации и устойчивости со сходимостью. При численном решении задач математической физики в общем случае необходимо исследовать и аппроксимацию, и устойчивость, и сходимость. Однако следующая теорема утверждает, что достаточно исследовать аппроксимацию и устойчивость, и, в случае положительного ответа, сходимость будет обеспечена.

Теорема эквивалентности. *Если конечно-разностная схема* (6.38) ***аппроксимирует*** *на точном решении дифференциальную задачу с p-м порядком по времени и q-м порядком по пространственной переменной и эта схема* ***устойчива****, то решение с помощью этой конечно-разностной схемы* ***сходится*** *к решению дифференциальной задачи с p-м порядком по времени и q-м порядком по пространственной переменной.*

Действительно, следующая цепочка доказывает теорему:

$$\|U - u_h\| \leqslant K \|f - f_h\| = K \|(LU)_h - L_h U_h\| = O\left(\tau^p + h^q\right),$$

откуда

$$\|U - u_h\| = O\left(\tau^p + h^q\right),$$

что и требовалась доказать.

В приведенной выше цепочке первое неравенство записано по условию устойчивости, а последнее равенство — по условию аппроксимации.

6.3.5. Консервативность и корректность. Все дифференциальные, интегральные и прочие уравнения характеризуют некоторые *законы сохранения* какой-либо субстанции (массы, энергии, импульса и т. п.). Заменяя дифференциальную задачу конечно-разностной схемой, можно нарушить эти законы сохранения.

Определение. *Конечно-разностная схема* ***консервативна****, если для нее выполняются законы сохранения, на основе которых поставлена дифференциальная задача.*

В противном случае конечно-разностная схема является ***неконсервативной***, т. е. решение с ее помощью не соответствует решению дифференциальной задачи (решается другая задача). Поэтому неконсервативными схемами пользоваться не рекомендуется.

Свойство *корректности* конечно-разностных схем вытекает из свойств аппроксимации и устойчивости.

Определение. *Дифференциальная задача* ***поставлена корректно****, если выполнены следующие два условия*:

1) *задача однозначно разрешима при любых входных данных*;

2) *решение задачи непрерывно зависит от входных данных.*

По аналогии определяется и корректность разностной задачи.

Определение. *Конечно-разностная задача* (6.38) ***поставлена корректно****, если при любых достаточно малых шагах* τ *и* h *сетки выполнены условия*:

1) *решение* u_h *существует и единственно при всех входных данных из некоторого допустимого семейства*;

2) *решение* u_h *непрерывно зависит от входных данных* f_h*, причем эта зависимость* **равномерна** *относительно величины шагов сетки* (*т.е. конечно-разностная схема устойчива*).

Таким образом, основными характеристиками конечно-разностной схемы, которые *обязательно* должны быть проанализированы, являются: *аппроксимация*, *устойчивость и консервативность*.

§ 6.4. Анализ порядка аппроксимации разностных схем

Из определения порядка аппроксимации ясно, что чем выше порядок аппроксимации, тем лучше конечно-разностная схема приближается к дифференциальной задаче. Это не означает, что решение по разностной схеме может быть так же близко к решению дифференциальной задачи, так как разностная схема может быть условно устойчивой или абсолютно неустойчивой вовсе.

Для нахождения порядка аппроксимации используется аппарат разложения в ряды Тейлора *точных* (неизвестных, но дифференцируемых) решений дифференциальной задачи в узлах сетки (подчеркнем: значения сеточной функции u_h дискретны, следовательно, не дифференцируемы и поэтому не разлагаются в ряды Тейлора).

В соответствии с определением порядка аппроксимации проанализируем порядок аппроксимации конечно-разностной схемы (6.28), для чего эту схему запишем на точном решении U_j^k:

$$\frac{U_j^{k+1} - U_j^k}{\tau} = a^2 \frac{U_{j+1}^k - 2U_j^k + U_{j-1}^k}{h^2}. \tag{6.45}$$

Разложим в ряды Тейлора по переменной x значения U_{j+1}^k, U_{j-1}^k в окрестности узла (j,k) до четвертой производной включительно, а значение U_j^{k+1} — в ряд Тейлора по переменной t в окрестности узла (j,k) до второй производной включительно, получим

$$U_{j+1}^k = U\left(x_j + h, t^k\right) = \\ = U_j^k + \left.\frac{\partial U}{\partial x}\right|_j^k h + \left.\frac{\partial^2 U}{\partial x^2}\right|_j^k \frac{h^2}{2} + \left.\frac{\partial^3 U}{\partial x^3}\right|_j^k \frac{h^3}{6} + O\left(h^4\right), \quad (6.46)$$

$$U_{j-1}^k = U\left(x_j - h, t^k\right) = \\ = U_j^k - \left.\frac{\partial U}{\partial x}\right|_j^k h + \left.\frac{\partial^2 U}{\partial x^2}\right|_j^k \frac{h^2}{2} - \left.\frac{\partial^3 U}{\partial x^3}\right|_j^k \frac{h^3}{6} + O\left(h^4\right), \quad (6.47)$$

$$U_j^{k+1} = U\left(x_j, t^k + \tau\right) = U_j^k + \left.\frac{\partial U}{\partial t}\right|_j^k \tau + O\left(\tau^2\right). \quad (6.48)$$

Подставляя (6.46)–(6.48) в (6.45), находим

$$L_h U_h = \frac{U_j^{k+1} - U_j^k}{\tau} - a^2 \frac{U_{j+1}^k - 2U_j^k + U_{j-1}^k}{h^2} = \\ = \left(\frac{\partial U}{\partial t} - a^2 \frac{\partial^2 U}{\partial x^2}\right)_j^k + O\left(\tau + h^2\right) = (LU)_h + O\left(\tau + h^2\right).$$

Таким образом,

$$\|(LU)_h - L_h U_h\| = O\left(\tau + h^2\right),$$

т. е. явная схема (6.28) для уравнения теплопроводности имеет *первый порядок аппроксимации по времени* и *второй — по пространственной переменной*. Аналогично, тот же порядок аппроксимации можно получить и для неявной схемы (6.30).

§ 6.5. Исследование устойчивости конечно-разностных схем

Поскольку устойчивость является одной из основных характеристик конечно-разностных схем, то в данном параграфе рассматриваются различные методы исследования устойчивости

конечно-разностных схем по начальным условиям. Наиболее распространенными методами исследования устойчивости являются следующие [10–14]:

— метод гармонического анализа (Фурье);

— принцип максимума;

— спектральный метод;

— энергетический метод.

Каждый из этих методов имеет достоинства и недостатки.

6.5.1. Метод гармонического анализа. Из математической физики известно, что решение начально-краевых задач представляется в виде следующего ряда:

$$u(x,t) = \sum_{n=1}^{\infty} u_n e^{i\lambda_n x},$$

где λ_n — собственные значения, а $\exp(i\lambda_n x) = \cos(\lambda_n x) + i\sin(\lambda_n x)$ — собственные функции, получаемые из решения соответствующей задачи Штурма–Лиувиля, т. е. решение может быть представлено в виде суперпозиции отдельных гармоник $u(x,t) = u_n(t)\exp(i\lambda_n x)$, каждая из которых есть произведение функции времени t и функции пространственной переменной x, причем последняя по модулю ограничена сверху единицей при любых значениях переменной x.

В то же время функция времени $u_n(t)$, называемая амплитудной частью гармоники, никак не ограничена, и, по всей вероятности, именно амплитудная часть гармоник является источником неконтролируемого входными данными роста функции и, следовательно, источником неустойчивости.

Таким образом, если конечно-разностная схема устойчива, то отношение амплитудной части гармоники на верхнем временном слое к амплитудной части на нижнем временном слое по модулю должно быть меньше единицы.

Если разложить значение сеточной функции u_j^k в ряд Фурье по собственным функциям:

$$u_j^k = \sum_{n=1}^{\infty} \eta_{kn} e^{i\lambda_n x_j}, \tag{6.49}$$

где амплитудная часть η_{kn} может быть представлена в виде произведения

$$\eta_{kn} = U_n \cdot q^k, \tag{6.50}$$

U_n — размерный и постоянный сомножитель амплитудной части, а k — показатель степени (соответствующий номеру временного слоя) сомножителя, зависящего от времени, то, подставив (6.49) в конечно-разностную схему, можно по модулю оценить отношение амплитудных частей на соседних временных слоях.

Однако поскольку операция суммирования линейна и собственные функции ортогональны для различных индексов суммирования, то в конечно-разностную схему вместо сеточных значений достаточно подставить одну гармонику разложения (6.49) (при этом у амплитудной части убрать индекс n), т. е.

$$u^k_{j\pm1} = \eta_k e^{i\lambda_n(x_j\pm h)}, \quad u^k_j = \eta_k e^{i\lambda_n(x_j)}, \quad u^{k+1}_j = \eta_{k+1} e^{i\lambda_n(x_j)}. \tag{6.51}$$

Таким образом, если конечно-разностная схема *устойчива по начальным данным*, то

$$\left|\frac{\eta_{k+1}}{\eta_k}\right| \leqslant 1, \tag{6.52}$$

т. е. условие (6.52) является *необходимым* условием устойчивости.

6.5.2. Исследование устойчивости методом гармонического анализа явной и неявной схем для уравнения теплопроводности. Подставим выражения (6.51) в явную конечно-разносную схему (6.28) для уравнения теплопроводности, получим

$$\frac{\eta_{k+1}}{\eta_k} = 1 - 4\sigma\cdot\sin^2\frac{\lambda_n h}{2}, \quad \sigma = \frac{a^2\tau}{h^2}. \tag{6.53}$$

Здесь использована формула, вытекающая из формулы Эйлера:

$$\exp(i\lambda_n h) - 2 + \exp(-i\lambda_n h) = -4\sin^2\frac{\lambda_n h}{2},$$

и формула $1 - \cos\lambda_n h = 2\sin^2\frac{\lambda_n h}{2}$, причем $0 < \sin^2\frac{\lambda_n h}{2} \leqslant 1$, поскольку $\lambda_n \neq 0$ и $h \neq 0$.

В соответствии с (6.53) получаем выражение

$$\left|\frac{\eta_{k+1}}{\eta_k}\right| = \left|1 - 4\sigma\sin^2\frac{\lambda_n h}{2}\right| = |1 - 4\sigma m|, \quad m = \sin^2\frac{\lambda_n h}{2},$$

или, с учетом (6.52), неравенство

$$|1 - 4\sigma m| \leqslant 1.$$

Отсюда получаем следующие два неравенства:

$$-1 \leqslant 1 - 4\sigma m \leqslant +1,$$

из которых правое выполнено всегда, а из левого следует знаменитое *условие устойчивости Куранта*: $\sigma m \leqslant 1/2$, или более жесткое для σ условие

$$\sigma = \frac{a^2\tau}{h^2} \leqslant \frac{1}{2}. \tag{6.54}$$

Из (6.54) следует, что *явная схема* для уравнения теплопроводности *условно устойчива* с условием (6.54), накладываемым на сеточные характеристики τ и h.

Подставим теперь гармоники (6.51) в неявную конечно-разностную схему (6.30) для уравнения теплопроводности, получим

$$\eta_{k+1} = \left(1 + 4\sigma \sin^2 \frac{\lambda_n h}{2}\right)^{-1} \eta_k,$$

откуда

$$\left|\frac{\eta_{k+1}}{\eta_k}\right| = \left|\frac{1}{1 + 4\sigma \sin^2 \dfrac{\lambda_n h}{2}}\right| \leqslant 1$$

всегда, так как σ и квадрат синуса больше нуля.

Следовательно, *неявная схема* для уравнения теплопроводности *абсолютно устойчива*, так как для выполнения неравенства $\left|\dfrac{\eta_{k+1}}{\eta_k}\right| \leqslant 1$ на сеточные характеристики τ и h не накладывалось никаких ограничений.

Комплекс $\dfrac{a^2\tau}{h^2}$ называют числом Куранта для уравнения теплопроводности.

6.5.3. Исследование устойчивости методом гармонического анализа явной и неявной схем для волнового уравнения. Конечно-разностные схемы (6.31) и (6.32) для волнового уравнения являются трехслойными, а не двухслойными, как для уравнения теплопроводности, поэтому подстановка в эти схемы гармоник в форме (6.51) не приведет к неравенству (6.52).

При исследовании устойчивости конечно-разностных схем для волнового уравнения методом гармонического анализа в конечно-разностные схемы подставляется отдельная гармоника ряда (6.49) с амплитудной частью в форме (6.50), т. е. гармоники вида

$$u_{j\pm1}^k = U_n \cdot q^k \cdot e^{i\lambda_n(x_j \pm h)}, \quad u_j^k = U_n \cdot q^k \cdot e^{i\lambda_n x_j},$$
$$u_j^{k\pm1} = U_n \cdot q^{k\pm1} \cdot e^{i\lambda_n x_j}. \tag{6.55}$$

Подставляем (6.55) в явную схему (6.31), получаем квадратное уравнение относительно параметра q:

$$q^2 - q\,(2 - 4\sigma m) + 1 = 0, \quad \sigma = \frac{a^2\tau^2}{h^2}, \quad m = \sin^2\frac{\lambda_n h}{2} > 0,$$

с корнями

$$q_{1,2} = (1 - 2\sigma m) \pm \sqrt{(1 - 2\sigma m)^2 - 1}\,.$$

Комплекс $\sigma = \dfrac{a^2\tau^2}{h^2}$ называют числом Куранта для волнового уравнения.

Таким образом, если явная схема (6.31) устойчива, то в соответствии с (6.50) и (6.52) величина q по модулю не превышает единицы

$$\left|(1 - 2\sigma m) \pm \sqrt{(1 - 2\sigma m)^2 - 1}\right| \leqslant 1, \tag{6.56}$$

или

$$-1 \leqslant (1 - 2\sigma m) \pm \sqrt{(1 - 2\sigma m)^2 - 1} \leqslant +1. \tag{6.57}$$

Здесь правое неравенство выполняется всегда.

Из левого неравенства (6.57) имеем

$$\pm\sqrt{(\sigma m - 1)\,\sigma m} \geqslant \sigma m - 1. \tag{6.58}$$

Для положительного знака в левой части (6.58) получаем, что $\sigma m > 1$. Действительно, в этом случае выполняется следующая цепочка неравенств: $(\sigma m - 1)\sigma m \geqslant (\sigma m - 1)^2$, $(\sigma m - 1)(\sigma m - \sigma m + 1) \geqslant 0$. Однако при этом не выполняется исходное неравенство (6.56), в чем можно убедиться непосредственной подстановкой $\sigma m \geqslant 1$ в (6.56).

Для отрицательного знака в левой части (6.58) имеем: $\sigma m \leqslant 1$, поскольку выполняются неравенства: $\sqrt{(\sigma m - 1)\sigma m} \leqslant \leqslant 1 - \sigma m$, $(\sigma m - 1)\sigma m \leqslant (1 - \sigma m)^2$, $(1 - \sigma m)(1 - \sigma m + + \sigma m) \geqslant 0$, $\sigma m \leqslant 1$. Это неравенство не противоречит неравенству (6.56), а поскольку $0 < m < 1$, то вместо неравенства $\sigma m \leqslant 1$ принимается более жесткое неравенство $\sigma \leqslant 1$ или

$$\sigma = \frac{a^2\tau^2}{h^2} \leqslant 1. \tag{6.59}$$

Таким образом, *явная конечно-разностная схема* (6.31) *для волнового уравнения условно устойчива с условием* (6.59), накладываемым на сеточные характеристики τ и h.

Для неявной схемы (6.32) результатом подстановки в нее гармоник (6.55) с амплитудой в форме (6.50) является квадратное уравнение

$$q^2\,(1 + 4\sigma m) - 2q + 1 = 0,$$

решением которого являются комплексные числа

$$q_{1,2} = \frac{1}{1 + 4\sigma m} \pm i\frac{2\sqrt{\sigma m}}{1 + 4\sigma m}.$$

Модуль этих комплексных чисел всегда не больше единицы:

$$|q_{1,2}| = \frac{\sqrt{1 + 4\sigma m}}{1 + 4\sigma m} \leqslant 1,$$

что доказывает *абсолютную устойчивость неявной схемы* (6.32) *для волнового уравнения.*

6.5.4. Принцип максимума. В математической физике известен принцип, в соответствии с которым решение начально-краевой задачи внутри расчетной области не может превышать значений искомой функции на пространственно-временной границе. Этот принцип положен в основу метода исследования устойчивости конечно-разностных схем, называемого *принципом максимума.*

Для его использования рассмотрим явную конечно-разностную схему (6.28) для уравнения теплопроводности в форме

$$u_j^{k+1} = \sigma u_{j+1}^k + (1 - 2\sigma)\, u_j^k + \sigma u_{j-1}^k \tag{6.60}$$

и введем норму сеточной функции u_j^k в виде $\left\|u^k\right\| = \max\limits_j \left|u_j^k\right|$. Тогда из (6.60) получим

$$\left\|u^{k+1}\right\| = \max_j \left|\sigma u_{j+1}^k + (1-2\sigma)\, u_j^k + \sigma u_{j-1}^k\right| \leqslant \leqslant \max_j \left|u_j^k\right| = \left\|u^k\right\|. \quad (6.61)$$

Если

$$\sigma \leqslant \frac{1}{2}, \quad (6.62)$$

то из (6.61) имеем неравенство

$$\left\|u^{k+1}\right\| \leqslant \left\|u^k\right\|,$$

откуда, продолжая цепочку неравенств вплоть до начального условия, получим

$$\left\|u^{k+1}\right\| \leqslant \left\|u^k\right\| \leqslant \ldots \leqslant \left\|u^0\right\| = \max_j \left|\psi(x_j)\right|, \quad (6.63)$$

где $\psi(x)$ — начальное условие (6.4).

Неравенства (6.63) в вычислительной математике называют *принципом максимума*. Он является достаточным условием устойчивости явной схемы (6.60) для уравнения теплопроводности.

Таким образом, если выполнено условие Куранта (6.62), то из цепочки (6.63) видно, что значение сеточной функции на любом временном слое по норме не превысит начального условия, т. е. рассматриваемая схема устойчива по начальному условию, причем условие (6.62) является теперь не только *необходимым* в соответствии с методом гармонического анализа, но и *достаточным*.

6.5.5. Спектральный метод исследования устойчивости. Рассмотрим сеточные функции u_j^k и u_j^{k+1}, $j = \overline{0, N}$, на двух временных слоях $t^k = k\tau$ и $t^{k+1} = (k+1)\tau$ и представим конечно-разностную схему в следующей операторной форме:

$$u^{k+1} = S \cdot u^k, \quad (6.64)$$

где S — оператор перехода от слоя t^k к слою t^{k+1}. Такой оператор можно построить не для всякой конечно-разностной схемы

(например, метод прогонки нельзя представить в форме (6.64)). Для явных конечно-разностных схем (6.28) или (6.31) оператор S представляется следующей матрицей перехода:

$$S = \begin{pmatrix} 1-2\sigma & \sigma & 0 & & \\ \sigma & 1-2\sigma & \sigma & & \\ & \sigma & 1-2\sigma & \sigma & \\ \dots & \dots & \dots & \dots & \dots \\ & & & \sigma & 1-2\sigma \end{pmatrix}.$$

Составим от левой и правой частей равенства (6.64) операцию нормы и используем свойство нормы: норма произведения операторов не превышает произведения норм, получим

$$\left\| u^{k+1} \right\| = \left\| S u^k \right\| \leqslant \|S\| \cdot \left\| u^k \right\|. \tag{6.65}$$

Если выполнено неравенство вида

$$\|S\| \leqslant 1, \tag{6.66}$$

то из условий (6.65) и (6.66) следует принцип максимума

$$\left\| u^{k+1} \right\| \leqslant \left\| u^k \right\| \leqslant \left\| u^{k-1} \right\| \leqslant \ldots \leqslant \left\| u^0 \right\| = \|\psi(x)\|.$$

Таким образом, если схема устойчива, то норма оператора перехода S не превышает единицы и, следовательно, условие (6.66) является *необходимым* условием устойчивости конечно-разностных схем.

Почему этот метод называют спектральным?

Для ответа на этот вопрос рассмотрим векторы u^k и u^{k+1} на соседних временных слоях в виде произведения сомножителя q^k, зависящего от времени (здесь k — показатель степени), на вектор $v(x)$, зависящий от пространственной переменной (для линейных задач метод разделения переменных легко обосновывается), получим равенства

$$u^k = q^k \cdot v, \quad u^{k+1} = q^{k+1} \cdot v,$$

подставляя которые в (6.64), будем иметь

$$Sv = qv. \tag{6.67}$$

Равенство (6.67) является задачей на собственные векторы v и собственные значения q оператора перехода S.

Составим от (6.67) операцию нормы, получим

$$\|qv\| = \|Sv\| \leqslant \|S\| \cdot \|v\|,$$

откуда

$$|q| \leqslant \|S\|, \tag{6.68}$$

Но для необходимого условия устойчивости выполняется неравенство (6.66), т. е. из (6.66) и (6.68) следует неравенство

$$|q| \leqslant 1. \tag{6.69}$$

Таким образом, если схема устойчива, то выполняется неравенство (6.69), называемое *условием устойчивости* фон Неймана, из которого ясно, что в этом случае все собственные значения q_j (или полный спектр оператора S) находятся внутри круга радиуса единица с центром в начале координат на комплексной плоскости.

6.5.6. Энергетический метод исследования устойчивости конечно-разностных схем. Как видно из предыдущих разделов, метод гармонического анализа и спектральный метод являются *необходимыми условиями устойчивости* конечно-разностных схем, а принцип максимума — *достаточным условием устойчивости.* В данном пункте рассматривается один из самых мощных и распространенных методов — энергетический метод, развитый в работах А. А. Самарского и базирующийся на понятиях энергетического пространства с энергетической нормой, энергетического тождества (неравенства) и принципа максимума. Ниже будет показано, что условия, используемые в энергетическом методе, являются *достаточными* условиями устойчивости конечно-разностных схем.

Для понимания энергетического метода рассмотрим применение его с целью исследования устойчивости конечно-разностных схем при численном решении следующей первой начально-краевой задачи для уравнения теплопроводности с однородными краевыми условиями:

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}, \quad x \in (0;1), \quad t > 0; \tag{6.70}$$

$$u(0,t)=0,\quad x=0,\quad t>0; \tag{6.71}$$

$$u(1,t)=0,\quad x=1,\quad t>0; \tag{6.72}$$

$$u(x,0)=\psi(x),\quad x\in[0;1],\quad t=0. \tag{6.73}$$

На сетке (6.25) будем аппроксимировать эту задачу с помощью явной (6.28) и неявной (6.30) конечно-разностных схем, записанных в векторно-операторной форме следующим образом:

$$\frac{u^{k+1}-u^k}{\tau}=\Lambda u^k, \tag{6.74}$$

$$\frac{u^{k+1}-u^k}{\tau}=\Lambda u^{k+1}, \tag{6.75}$$

где конечно-разностный оператор Λ аппроксимирует дифференциальный оператор по пространственной переменной x, т. е.

$$\Lambda u^k=a^2(u^k_{j+1}-2u^k_j+u^k_{j-1})/h^2.$$

Энергетическое пространство. Введем энергетическое пространство H_A сеточных функций u_h, являющееся гильбертовым пространством, в котором определено скалярное произведение для двух элементов $u_h\in H_A$ и $v_h\in H_A$:

$$(u_h,v_h)_A=(Au_h,v_h)=\sum_{j=1}^{N}Au_jv_jh, \tag{6.76}$$

и, следовательно, с нормой

$$\|u_h\|_A=(Au_h,u_h)^{1/2}=\left[\sum_{j=1}^{N}Au_ju_jh\right]^{1/2}. \tag{6.77}$$

Как известно, гильбертово пространство — это полное нормированное пространство, в котором определено скалярное произведение. Здесь *полнота* определяется в том смысле, что если последовательность сеточных функций $\{u_{h_n}\}$ сходится к своему пределу при $h_n\to 0$ (в данном случае — к решению дифференциальной задачи), то она является *фундаментальной*, т. е. выполняется *условие Коши*

$$\|u_{h_n}-u_{h_m}\|\xrightarrow[n,m\to\infty]{}0. \tag{6.78}$$

Действительно, если конечно-разностная схема аппроксимирует дифференциальную задачу и устойчива, то по теореме эквивалентности решение с помощью конечно-разностной схемы сходится к решению дифференциальной задачи при *измельчении сетки.*

Для двух сеточных функций u_{h_n} и u_{h_m} (двух элементов гильбертова пространства H_A) на различных сетках с шагами h_n и h_m понятие полноты означает, что при измельчении сетки, т. е. при $h_n \to 0$ (или $h_m \to 0$) последовательности $\{u_{h_n}\}$ и $\{u_{h_m}\}$ сходятся к одному и тому же пределу, т. е. выполняется (6.78).

В дальнейшем потребуются следующие понятия, характеризующие конечно-разностные операторы: сопряженность, самосопряженность, положительная определенность.

Определение. Конечно-разностный оператор A^* называется *сопряженным оператору* A, если выполняется равенство

$$(Au, v) = (u, A^*v). \tag{6.79}$$

Например, если оператор A — симметрическая матрица с действительными элементами ($A = A^T$), то A — сопряженный оператор (это можно проверить непосредственно).

Определение. Конечно-разностный оператор A называется *самосопряженным*, если выполняется равенство

$$(Au, v) = (u, Av). \tag{6.80}$$

Определение. Конечно-разностный оператор A называется *положительно определенным* ($A > 0$) или *положительно полуопределенным* ($A \geqslant 0$) на гильбертовом пространстве сеточных функций $u_h \in H_A$, если

$$(Au_h, u_h) > 0 \quad \text{или} \quad (Au_h, u_h) \geqslant 0. \tag{6.81}$$

Можно показать, что разностный оператор $-\Lambda = A$ является самосопряженным, т. е.

$$(-\Lambda u, v) = (u, \ -\Lambda v)\,.$$

С целью определения собственных функций и собственных значений конечно-разностного оператора A, рассмотрим вначале задачу на собственные значения и собственные функции оператора Λ:

$$\Lambda\varphi = \lambda\varphi. \tag{6.82}$$

Собственные функции $\varphi_n(x_j)$ должны удовлетворять следующим условиям:

1) быть ортогональными на отрезке $x \in [0;1]$ при $n \neq m$, т. е. $(\varphi_n, \varphi_m) = 0$ на $x \in [0;1]$ при $n \neq m$;

2) удовлетворять однородным краевым уравнениям (6.71), (6.72);

3) их число должно совпадать с числом собственных значений λ_n.

Таким условиям удовлетворяют функции:

$$\varphi_n(x_j) = \sin\frac{n\pi x_j}{l}, \quad x_j = jh, \quad l = 1, \quad h = \frac{1}{N},$$

т. е.

$$\varphi_n(x_j) = \sin\frac{n\pi j}{N}, \quad j = \overline{0, N}; \quad n = \overline{1, N-1}. \tag{6.83}$$

Для нахождения собственных значений λ_n подставим (6.83) в (6.82), получим

$$\Lambda\left(\sin\frac{n\pi j}{N}\right) = \lambda_n \sin\frac{n\pi j}{N},$$

или

$$\frac{a^2}{h^2}\left[\sin\frac{n\pi(j+1)}{N} - 2\sin\frac{n\pi j}{N} + \sin\frac{n\pi(j-1)}{N}\right] = \lambda_n \sin\frac{n\pi j}{N},$$

откуда

$$\lambda_n = -\frac{4a^2}{h^2}\sin^2\frac{n\pi}{2N}, \quad n = \overline{1, N-1}. \tag{6.84}$$

Из (6.84) видно, что все собственные значения оператора Λ отрицательны, а собственные значения оператора $A = -\Lambda$ положительны, т. е. оператор $A = -\Lambda$ *положительно определен.*

При исследовании устойчивости явной конечно-разностной схемы (6.74) *энергетическим методом* воспользуемся следующими тождествами:

$$\begin{gathered} u_t = \frac{u^{k+1} - u^k}{\tau}; \\ u^k = \frac{u^{k+1} + u^k}{2} - \frac{u^{k+1} - u^k}{2} = \frac{u^{k+1} + u^k}{2} - \frac{\tau}{2}u_t. \end{gathered} \tag{6.85}$$

Умножим скалярно явную схему (6.74) на вектор $2\tau u_t$, получим

$$2\tau\,(u_t, u_t) + \left(Au^k, 2\tau u_t\right) = 0,$$

или, после подстановки сюда тождеств (6.85),

$$2\tau\,(u_t, u_t) + \left(Au^{k+1}, u^{k+1}\right) - \left(Au^k, u^k\right) - \\ - \tau^2\,(Au_t, u_t) - \left(Au^{k+1}, u^k\right) + \left(Au^k, u^{k+1}\right) = 0.$$

В силу самосопряженности оператора A и коммутативности скалярного произведения, последние два слагаемых сокращаются, после чего получим следующее *энергетическое тождество*:

$$2\tau\left(\left(E - \frac{\tau}{2}A\right)u_t, u_t\right) + \left(Au^{k+1}, u^{k+1}\right) = \left(Au^k, u^k\right). \quad (6.86)$$

Если оператор

$$E - (\tau/2)\,A \geqslant 0, \quad (6.87)$$

то из (6.86) получаем следующее *энергетическое неравенство*:

$$\left(Au^{k+1}, u^{k+1}\right) \leqslant \left(Au^k, u^k\right), \quad (6.88)$$

или

$$\left\|u^{k+1}\right\|_A \leqslant \left\|u^k\right\|_A,$$

откуда сразу следует принцип максимума

$$\left\|u^{k+1}\right\|_A \leqslant \left\|u^k\right\|_A \leqslant \ldots \leqslant \left\|u^0\right\|_A = \left\|\psi\,(x)\right\|_A,$$

являющийся достаточным условием устойчивости.

Если теперь от неравенства (6.87) вычислить любую норму ($\|E\| = 1$), например норму, которая равна максимальному по модулю собственному значению оператора A, то с использованием выражения (6.84) получим

$$\frac{\tau}{2}\,\|A\|_C \leqslant \|E\|, \quad \|A\|_C = \max_n |\lambda_n| = \frac{4a^2}{h^2}; \quad \frac{4a^2}{h^2}\cdot\frac{\tau}{2} \leqslant 1,$$

откуда

$$\frac{a^2\tau}{h^2} \leqslant \frac{1}{2}. \quad (6.89)$$

Таким образом, условие устойчивости Куранта (6.54) явной конечно-разностной схемы для уравнения теплопроводности, выведенное с помощью метода гармонического анализа, является и *достаточным условием*.

Исследуем теперь устойчивость неявной конечно-разностной схемы (6.75) энергетическим методом, для чего к тождествам (6.85) добавим тождество

$$u^{k+1} = u^k + \tau \cdot u_t = \frac{u^{k+1} + u^k}{2} + \frac{\tau}{2} u_t. \qquad (6.90)$$

Умножим скалярно схему (6.75) на вектор $2\tau \cdot u_t$, получим

$$2\tau\, (u_t, u_t) + A\left(u^k + \tau \cdot u_t,\, 2\tau \cdot u_t\right) = 0,$$

$$2\tau \left(\left(E + \frac{\tau}{2} A\right) u_t, u_t\right) + 2\tau \left(A \frac{u^{k+1} + u^k}{2}, \frac{u^{k+1} - u^k}{\tau}\right) = 0,$$

$$2\tau \left(\left(E + \frac{\tau}{2} A\right) u_t, u_t\right) + \left(Au^{k+1}, u^{k+1}\right) - \left(Au^k, u^k\right) - \\ - \left(Au^{k+1}, u^k\right) + \left(Au^k, u^{k+1}\right) = 0.$$

Таким образом, для неявной схемы *энергетическое тождество* имеет вид

$$2\tau \left(\left(E + \frac{\tau}{2} A\right) u_t, u_t\right) + \left(Au^{k+1}, u^{k+1}\right) = \left(Au^k, u^k\right). \qquad (6.91)$$

Здесь первое слагаемое всегда положительно определено, поэтому *энергетическое неравенство* имеет вид

$$\left(Au^{k+1}, u^{k+1}\right) \leqslant \left(Au^k, u^k\right), \\ \left\|u^{k+1}\right\|_A \leqslant \left\|u^k\right\|_A, \qquad (6.92)$$

откуда следует принцип максимума

$$\left\|u^{k+1}\right\|_A \leqslant \left\|u^k\right\|_A \leqslant \ldots \leqslant \left\|u^0\right\|_A = \|\psi(x)\|_A.$$

Таким образом, *неявная схема* (6.75) *безусловно устойчива*, так как оператор $E + \dfrac{\tau}{2} A$ всегда положительно определен.

§ 6.6. Конечно-разностный метод решения задач для уравнений параболического типа

Конечно-разностные схемы и методы численного решения одномерных и многомерных по пространству начально-краевых задач для уравнений параболического типа, излагаемые в этом параграфе, можно использовать и для соответствующих задач для уравнений гиперболического типа, а некоторые — методы расщепления — и в задачах для уравнений эллиптического типа.

6.6.1. Однородные и консервативные конечно-разностные схемы для задач теплопроводности с граничными условиями, содержащими производные. В задачах математической физики вообще, и в задачах теплопроводности в частности, граничные условия 1-го рода аппроксимируются точно в узлах на границе расчетной области и этот факт никак не влияет на порядок аппроксимации во всей расчетной области. Этого нельзя сказать об аппроксимации граничных условий 2-го и 3-го родов, поскольку в них присутствует производная первого порядка искомой функции по пространственной переменной, в результате чего порядок аппроксимации в граничных узлах может быть ниже порядка аппроксимации во внутренних узлах расчетной области.

В этой связи необходимо ввести понятия *локального порядка* аппроксимации — порядка аппроксимации в отдельно взятом узле, и *глобального порядка* аппроксимации — порядка аппроксимации во всей расчетной области. Анализ порядка аппроксимации, рассмотренный в § 6.4 осуществлен для отдельного узла сетки, и, следовательно, там речь идет о локальной аппроксимации. Для задач с граничными условиями 1-го рода на *равномерной сетке* локальный порядок аппроксимации совпадает с глобальным.

Для задач с граничными условиями, содержащими производные, порядок аппроксимации в граничных узлах может быть ниже порядка аппроксимации во внутренних узлах, если конечно-разностная схема в граничных узлах не использует дифференциальное уравнение (ниже будет показано, что в этом случае нарушается консервативность). Тогда глобальный порядок аппроксимации равен *наименьшему* относительно всех узлов сетки порядку аппроксимации.

Рассмотрим методологию сохранения на границах с граничными условиями 2-го и 3-го родов порядка аппроксимации, ко-

торый дает аппроксимация дифференциального уравнения во внутренних узлах, т. е. получение конечно-разностной схемы с однородной аппроксимацией. Для этого рассмотрим третью начально-краевую задачу для уравнения теплопроводности, содержащего как конвективные члены (пропорциональные производной $\partial u/\partial x$), так и источниковые члены, содержащие искомую функцию $u(x,t)$:

$$\begin{cases} \dfrac{\partial u}{\partial t} = a^2\dfrac{\partial^2 u}{\partial x^2} + b\dfrac{\partial u}{\partial x} + cu, \quad 0 < x < l, \quad t > 0; & (6.93) \\[2ex] \alpha\dfrac{\partial u\,(0,t)}{\partial x} + \beta u\,(0,t) = \varphi_0\,(t)\,, \quad x = 0, \quad t > 0; & (6.94) \\[2ex] \gamma\dfrac{\partial u\,(l,t)}{\partial x} + \delta u\,(l,t) = \varphi_1\,(t)\,, \quad x = l, \quad t > 0; & (6.95) \\[2ex] u\,(x,0) = \psi\,(x)\,, \quad 0 \leqslant x \leqslant l, \quad t = 0. & (6.96) \end{cases}$$

Во внутренних узлах конечно-разностной сетки (6.25) неявная конечно-разностная схема для уравнения (6.93) имеет вид

$$\frac{u_j^{k+1} - u_j^k}{\tau} = \frac{a^2}{h^2}\left(u_{j+1}^{k+1} - 2u_j^{k+1} + u_{j-1}^{k+1}\right) + \frac{b}{2h}\left(u_{j+1}^{k+1} - u_{j-1}^{k+1}\right) + \\ + cu_j^{k+1} + O\left(\tau + h^2\right), \quad j = \overline{1, N-1}. \quad (6.97)$$

Если производные первого порядка в граничных условиях (6.94) и (6.95) аппроксимировать по следующей схеме (с помощью отношения конечных разностей справа и слева):

$$\left.\frac{\partial u}{\partial x}\right|_{j=0}^{k+1} = \frac{u_1^{k+1} - u_0^{k+1}}{h} + O\,(h)\,;$$

$$\left.\frac{\partial u}{\partial x}\right|_{j=N}^{k+1} = \frac{u_N^{k+1} - u_{N-1}^{k+1}}{h} + O\,(h)\,,$$

то граничные условия аппроксимируются с первым порядком, и глобальный порядок будет равен первому порядку, несмотря на то что во всех остальных узлах порядок аппроксимации по пространственной переменной равен двум. Для сохранения порядка аппроксимации, равного двум, в граничных узлах разложим *на точном решении* значение u_1^{k+1} в окрестности точки $x = 0$ в ряд

Тейлора по переменной x до третьей производной включительно, а u_{N-1}^{k+1} — в аналогичный ряд в окрестности точки $x = l$, получим (в предположении что функция $u(x,t)$ в граничных узлах имеет первые производные по времени и вторые — по x):

$$u_1^{k+1} = u\left(0 + h, t^{k+1}\right) = \\ = u_0^{k+1} + \left.\frac{\partial u}{\partial x}\right|_0^{k+1} h + \left.\frac{\partial^2 u}{\partial x^2}\right|_0^{k+1} \frac{h^2}{2} + O\left(h^3\right), \quad (6.98)$$

$$u_{N-1}^{k+1} = u\left(l - h, t^{k+1}\right) = \\ = u_N^{k+1} - \left.\frac{\partial u}{\partial x}\right|_N^{k+1} h + \left.\frac{\partial^2 u}{\partial x^2}\right|_N^{k+1} \frac{h^2}{2} + O\left(h^3\right). \quad (6.99)$$

Подставим сюда значения второй производной в граничных узлах, полученные из дифференциального уравнения (6.93):

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{j=0,N}^{k+1} = \left(\frac{1}{a^2} \cdot \frac{\partial u}{\partial t} - \frac{b}{a^2} \cdot \frac{\partial u}{\partial x} - \frac{c}{a^2} u\right)_{j=0,N}^{k+1},$$

и найдем из полученных выражений (6.98), (6.99) значения первой производной $\left.\frac{\partial u}{\partial x}\right|_{j=0,N}^{k+1}$ в граничных узлах с порядком $O\left(\tau + h^2\right)$:

$$\left.\frac{\partial u}{\partial x}\right|_0^{k+1} = \frac{2a^2}{h\left(2a^2 - bh\right)} \cdot \left(u_1^{k+1} - u_0^{k+1}\right) - \\ - \frac{h}{2a^2 - bh} \cdot \left.\frac{\partial u}{\partial t}\right|_0^{k+1} + \frac{ch}{2a^2 - bh} \cdot u_0^{k+1} + O\left(h^2\right),$$

$$\left.\frac{\partial u}{\partial x}\right|_N^{k+1} = \frac{2a^2}{h\left(2a^2 + bh\right)} \cdot \left(u_N^{k+1} - u_{N-1}^{k+1}\right) + \\ + \frac{h}{2a^2 + bh} \cdot \left.\frac{\partial u}{\partial t}\right|_N^{k+1} - \frac{ch}{2a^2 + bh} \cdot u_N^{k+1} + O\left(h^2\right).$$

Подставляя $\left.\dfrac{\partial u}{\partial x}\right|_0^{k+1}$ в (6.94), а $\left.\dfrac{\partial u}{\partial x}\right|_N^{k+1}$ в (6.95) и аппроксимируя полученные соотношения в соответствующих граничных узлах (при этом $\left.\dfrac{\partial u}{\partial t}\right|_0^{k+1} = \left(u_0^{k+1} - u_0^k\right)/\tau + O(\tau)$, $\left.\dfrac{\partial u}{\partial t}\right|_N^{k+1} = \left(u_N^{k+1} - u_N^k\right)/\tau + O(\tau)$), получим алгебраические уравнения для граничных узлов, в каждом из которых два неизвестных:

$$b_0 \cdot u_0^{k+1} - c_0 \cdot u_1^{k+1} = d_0 + O\left(\tau + h^2\right), \quad j = 0; \tag{6.100}$$

$$a_0 = 0, \quad b_0 = \frac{2a^2}{h} + \frac{h}{\tau} - ch - \frac{\beta}{\alpha}\left(2a^2 - bh\right); \quad c_0 = \frac{2a^2}{h};$$

$$d_0 = \frac{h}{\tau} \cdot u_0^k - \varphi_0\left(t^{k+1}\right) \cdot \frac{2a^2 - bh}{\alpha};$$

$$-a_N \cdot u_{N-1}^{k+1} + b_N \cdot u_N^{k+1} = d_N + O\left(\tau + h^2\right), \quad j = N; \tag{6.101}$$

$$a_N = \frac{2a^2}{h}; \quad b_N = \frac{2a^2}{h} + \frac{h}{\tau} - ch + \frac{\delta}{\gamma} \cdot \left(2a^2 + bh\right); \quad c_n = 0;$$

$$d_N = \frac{h}{\tau} \cdot u_N^k + \varphi_1\left(t^{k+1}\right) \cdot \frac{2a^2 + bh}{\gamma}.$$

Таким образом, (6.100) — конечно-разностная аппроксимация граничного условия 3-го рода (6.94) на левой границе $x = 0$, а (6.101) — конечно-разностная аппроксимация граничного условия 3-го рода (6.95) на правой границе $x = l$, которые сохраняют тот же порядок аппроксимации, что и в конечно-разностной аппроксимации (6.97) дифференциального уравнения (6.93).

Приписывая к граничным конечно-разностным уравнениям (6.100), (6.101), каждое из которых содержит два значения сеточной функции, алгебраические уравнения (6.97), записанные в виде

$$-a_j u_{j-1}^{k+1} + b_j u_j^{k+1} - c_j u_{j+1}^{k+1} = d_j + O\left(\tau + h^2\right), \quad j = \overline{1, N-1}; \tag{6.102}$$

$$a_j = \frac{a^2}{h^2} - \frac{b}{2h}; \quad b_j = \frac{2a^2}{h^2} + \frac{1}{\tau} - c; \quad c_j = \frac{a^2}{h^2} + \frac{b}{2h}; \quad d_j = \frac{1}{\tau} \cdot u_j^k,$$

получим СЛАУ с трехдиагональной матрицей, решаемую методом прогонки ($a_0 = 0$; $c_N = 0$)

$$A_j = \frac{c_j}{b_j - a_j A_{j-1}};$$

$$B_j = \frac{d_j + a_j B_{j-1}}{b_j - a_j A_{j-1}} \quad \left(A_0 = \frac{c_0}{b_0};\ B_0 = \frac{d_0}{b_0};\ A_N = 0\right), \quad j = \overline{0, N}; \tag{6.103}$$

$$u_j^{k+1} = A_j u_{j+1}^{k+1} + B_j \quad \left(u_N^{k+1} = B_N\right), \quad j = N, N-1, \ldots, 0. \tag{6.104}$$

Покажем, что изложенный метод аппроксимации краевых условий, содержащих производные по пространственным переменным, не только повышает порядок аппроксимации, но и сохраняет консервативность конечно-разностной схемы, т. е. в конечно-разностной аппроксимации соблюдаются законы сохранения, на основе которых выведены дифференциальные соотношения задачи (6.93)–(6.96). Для этого рассмотрим вначале вывод дифференциального уравнения теплопроводности (6.93) в случае $b = c = 0$ (см. рис. 6.6 *а*).

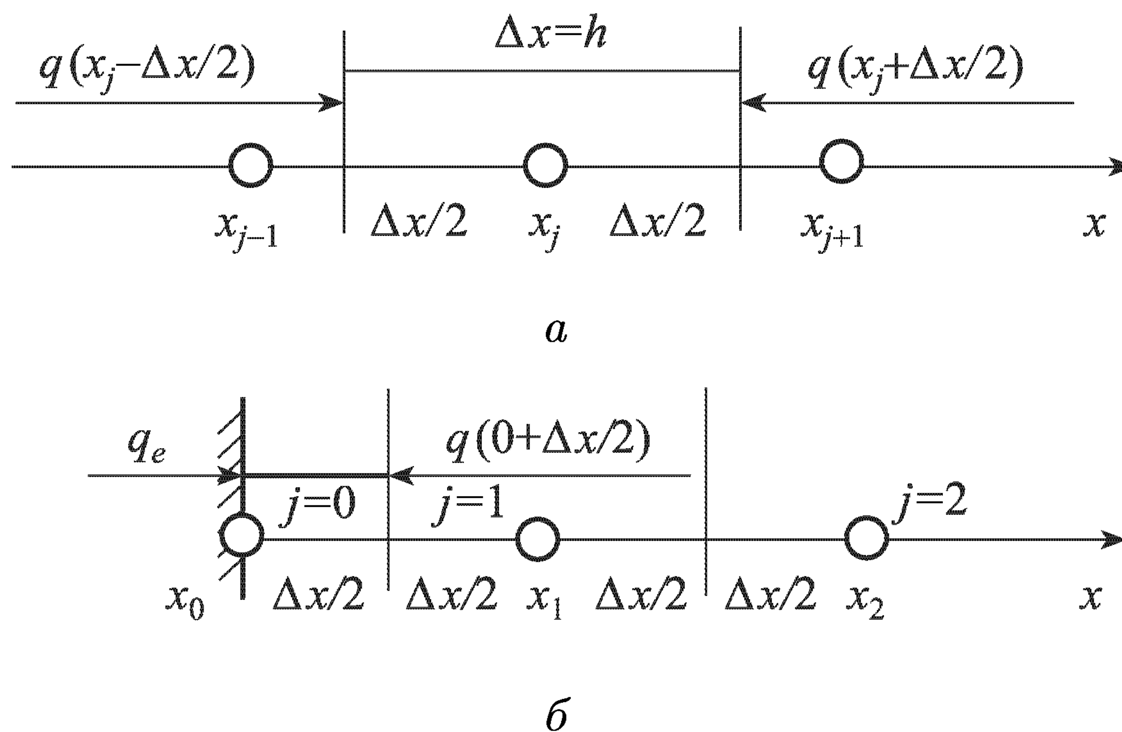


Рис. 6.6. К аппроксимации краевых условий, содержащих производные, с сохранением консервативности

Из физики известно, что тепловой поток, согласно закону Фурье, в одномерном случае равен $q = -\lambda \cdot \dfrac{\partial u}{\partial x}$, где λ — коэффициент теплопроводности. Для элемента длиной $\Delta x = h$, в центре которого помещен узел x_j, запишем сумму тепловых потоков:

$q\,(x_j - \Delta x/2) - q\,(x_j + \Delta x/2)$, подходящих к левой и правой границам элемента. По закону сохранения энергии сумма этих тепловых потоков равна изменению энергии в этом элементе Δx, которая пропорциональна массе элемента $\rho\Delta x$, теплоемкости материала c и производной первого порядка от температуры $u(x,t)$ по времени

$$-\lambda \cdot \frac{\partial u\,(x_j - \Delta x/2)}{\partial x} - \left(-\lambda \cdot \frac{\partial u\,(x_j + \Delta x/2)}{\partial x}\right) = c \cdot \rho \cdot \Delta x \cdot \frac{\partial u}{\partial t}.$$

Разделив это равенство на Δx и перейдя к пределу при $\Delta x \to 0$, получим одномерное уравнение теплопроводности

$$\frac{\partial u}{\partial t} = a^2 \cdot \frac{\partial^2 u}{\partial x^2}, \quad a^2 = \frac{\lambda}{c\rho}.$$

Таким образом, конечно-разностная аппроксимация (6.97) дифференциального уравнения (6.93) учитывает и тепловые потоки, и энергию, поглощенную элементом Δx, т. е. все виды энергии, участвующие при выводе дифференциального уравнения.

Рассмотрим теперь вывод граничного условия (6.94) на левой границе (см. рис. 6.6 *б*) расчетной области (для правой границы $x = l$ вывод аналогичен).

К граничному узлу $x = 0$ примыкает масса объемом $\Delta x/2$ со стороны расчетной области. При задании граничного условия 3-го рода на левой границе $x = 0$ осуществляется теплообмен с окружающей средой по закону Ньютона: $q_e = \beta\,(u_e - u\,(0,t))$, где u_e — температура окружающей среды, а β — коэффициент теплообмена на границе $x = 0$, имеющей температуру $u(0,t)$, с окружающей средой, имеющей температуру u_e. Справа к элементу $\Delta x/2$ подводится тепловой поток, описываемый законом Фурье. Тогда разность тепловых потоков, подводимых к половине элемента, равна энергии, пошедшей на повышение температуры элемента $\Delta x/2$, пропорциональной массе этого элемента $\rho\Delta x/2$, теплоемкости c и производной температуры по времени:

$$\beta\,(u_e - u\,(0,t)) - \left(-\lambda \cdot \frac{\partial u\,(0 + \Delta x/2)}{\partial x}\right) = c\rho \cdot \frac{\Delta x}{2} \cdot \frac{\partial u}{\partial t}. \quad (6.105)$$

Переходя здесь к пределу при $\Delta x/2 \to 0$, получим левое граничное условие 3-го рода (6.94) (с точностью до коэффициентов)

$$\lambda \cdot \frac{\partial u(0,t)}{\partial x} + \beta \left(u_e - u(0,t)\right) = 0. \tag{6.106}$$

Таким образом, конечно-разностная аппроксимация граничного условия (6.106) или (6.94) должна сопровождаться появлением *консервативного слагаемого*, стоящего в правой части равенства (6.105). Физически это означает, что граничные условия (6.94), (6.95) записаны для границ, не имеющих ни массы, ни объема. В конечно-разностной аппроксимации к границе примыкает масса с одномерным объемом, равным $\Delta x/2$, в котором необходимо учитывать дифференциальное уравнение, действующее во внутренних точках расчетной области, и не действующее на границе.

В конечно-разностной аппроксимации краевого условия, содержащего производную по переменной x, нестационарный член в правой части выражения (6.105) с помощью дифференциального уравнения можно заменить на дивергентный член, пропорциональный $\partial^2 u/\partial x^2\,|_{x=0}$.

Аналогичный подход можно осуществить в краевых задачах для дифференциальных уравнений любых типов.

6.6.2. Неявно-явная конечно-разностная схема с весами. Схема Кранка–Николсона. Явная конечно-разностная схема (6.28), записанная в форме

$$u_j^{k+1} = \sigma \cdot u_{j+1}^k + (1 - 2\sigma)\, u_j^k + \sigma \cdot u_{j-1}^k, \quad \sigma = \frac{a^2 \tau}{h^2},$$

$$j = \overline{1, N-1}, \quad k = 0, 1, 2 \ldots, \tag{6.107}$$

обладает тем *достоинством*, что решение на верхнем временном слое t^{k+1} получается сразу (без решения СЛАУ) по значениям сеточной функции на нижнем временном слое t^k, где решение известно (при $k = 0$ значения сеточной функции формируются из начального условия (6.4.)). Но эта же схема обладает существенным *недостатком*, поскольку она является условно устойчивой с условием (6.54), накладываемым на сеточные характеристики τ и h.

С другой стороны, неявная конечно-разностная схема (6.30), записанная форме

$$-a_j u_{j-1}^{k+1} + b_j u_j^{k+1} - c_j u_{j+1}^{k+1} = d_j, \;\; j = \overline{1, N-1}, \;\; k = 0, 1, 2, \ldots, \tag{6.108}$$

приводит к необходимости решать СЛАУ, но зато эта схема абсолютно устойчива.

Проанализируем схемы (6.107), (6.108). Пусть точное решение, которое неизвестно, возрастает по времени, т. е. $u_j^{k+1} > u_j^k$. Тогда, в соответствии с явной схемой (6.107), разностное решение будет заниженным по сравнению с точным, так как u_j^{k+1} определяется по меньшим значениям сеточной функции на предыдущем временном слое, поскольку решение является возрастающим по времени.

Для неявной схемы (6.108) на возрастающем решении, наоборот, решение завышено по сравнению с точным, поскольку оно определяется по значениям сеточной функции на верхнем временном слое.

На убывающем решении картина изменяется противоположным образом: явная конечно-разностная схема завышает решения, а неявная — занижает (см. рис. 6.7).

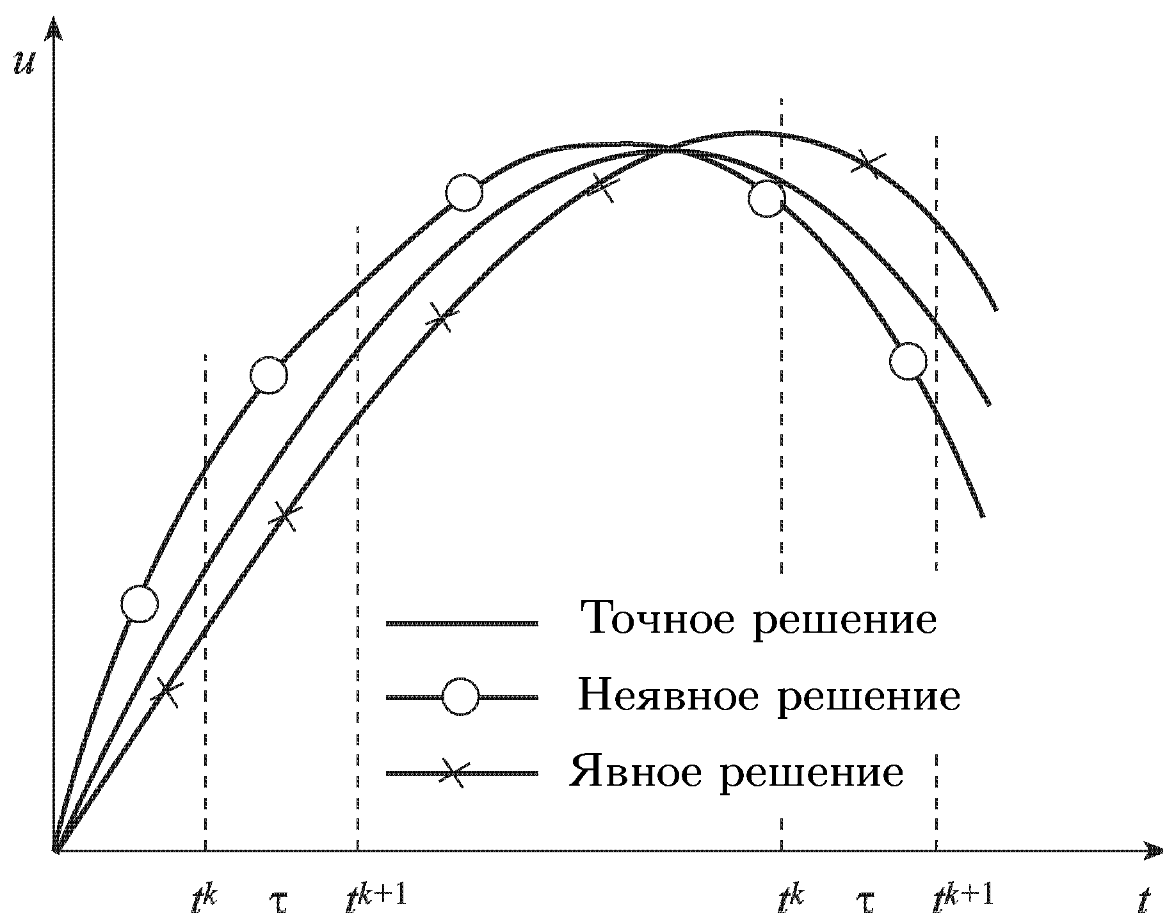


Рис. 6.7. Двусторонний метод аппроксимации

На основе этого анализа возникла идея о построении более точной неявно-явной конечно-разностной схемы с весами при пространственных конечно-разностных операторах, причем при измельчении шагов τ и h точное (неизвестное) решение может быть взято в «вилку» сколь угодно узкую, так как если явная и неявная схемы аппроксимируют дифференциальную задачу и эти схемы устойчивы, то при стремлении сеточных характери-

стик τ и h к нулю решения по явной и неявной схемам стремятся к точному решению с разных сторон.

Проведенный анализ дал блестящий пример так называемых двусторонних методов, исследованных *В. К. Саульевым* [15].

Рассмотрим неявно-явную схему с весами для простейшего уравнения теплопроводности:

$$\frac{u_j^{k+1} - u_j^k}{\tau} = \theta a^2 \frac{u_{j+1}^{k+1} - 2u_j^{k+1} + u_{j-1}^{k+1}}{h^2} + \\ + (1-\theta)\, a^2 \frac{u_{j+1}^k - 2u_j^k + u_{j-1}^k}{h^2}, \quad (6.109)$$

где θ — вес неявной части конечно-разностной схемы, $1-\theta$ — вес для явной части, причем $0 \leqslant \theta \leqslant 1$. При $\theta = 1$ имеем полностью неявную схему, при $\theta = 0$ — полностью явную схему, и при $\theta = 1/2$ — схему *Кранка–Николсона.*

В соответствии с гармоническим анализом для схемы (6.109) получаем неравенство

$$\left|\frac{\eta_{k+1}}{\eta_k}\right| < \left|\frac{1 - 4\sigma(1-\theta)}{1 + 4\sigma\theta}\right| \leqslant 1,$$

откуда

$$-1 \leqslant \frac{1 - 4\sigma(1-\theta)}{1 + 4\sigma\theta} \leqslant +1, \quad (6.110)$$

причем правое неравенство выполнено всегда.

Левое неравенство имеет место для любых значений σ, если $1/2 \leqslant \theta \leqslant 1$. Если же вес θ лежит в пределах $0 \leqslant \theta < 1/2$, то между σ и θ из левого неравенства устанавливается связь

$$\sigma \leqslant \frac{1}{2\cdot(1-2\theta)}, \quad 0 \leqslant \theta < 1/2, \quad (6.111)$$

являющаяся условием устойчивости неявно-явной схемы с весами (6.109), когда вес находится в пределах $0 \leqslant \theta < 1/2$.

Таким образом, неявно-явная схема с весами (6.109) абсолютно устойчива при $1/2 \leqslant \theta \leqslant 1$ и условно устойчива с условием (6.111) при $0 \leqslant \theta < 1/2$.

Рассмотрим порядок аппроксимации неявно-явной схемы с весами (6.109), для чего разложим в ряд Тейлора в окрестности

узла (x_j, t^k) на точном решении значения сеточных функций u_j^{k+1} по переменной t, $u_{j\pm1}^k$, $u_{j\pm1}^{k+1}$ по переменной x и полученные разложения подставим в (6.109):

$$\left.\frac{\partial u}{\partial t}\right|_j^k + \left.\frac{\partial^2 u}{\partial t^2}\right|_j^k \frac{\tau}{2} + O\left(\tau^2\right) =$$

$$= \theta \cdot a^2 \frac{\partial^2}{\partial x^2}\left[u_j^k + \left.\frac{\partial u}{\partial t}\right|_j^k \cdot \tau + O\left(\tau^2\right)\right] +$$

$$+ (1-\theta)\, a^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_j^k + O\left(h^2\right).$$

В этом выражении дифференциальный оператор $\frac{\partial^2}{\partial x^2}$ от квадратной скобки в соответствии с дифференциальным уравнением (6.1) равен дифференциальному оператору $\frac{1}{a^2}\frac{\partial}{\partial t}$, в соответствии с чем вышеприведенное равенство приобретает вид

$$\left.\frac{\partial u}{\partial t}\right|_j^k + \left.\frac{\partial^2 u}{\partial t^2}\right|_j^k \frac{\tau}{2} + O\left(\tau^2\right) =$$

$$= \theta a^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_j^k + \theta \left.\frac{\partial^2 u}{\partial t^2}\right|_j^k \cdot \tau + a^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_j^k -$$

$$- \theta a^2 \left.\frac{\partial^2 u}{\partial x^2}\right|_j^k + O\left(\tau^2 + h^2\right).$$

После упрощения получаем

$$\left(\frac{\partial u}{\partial t} - a^2 \frac{\partial^2 u}{\partial x^2}\right)_j^k = \left(\theta - {}^1\!/_2\right) \left.\frac{\partial^2 u}{\partial t^2}\right|_j^k \cdot \tau + O\left(\tau^2 + h^2\right),$$

откуда видно, что для схемы Кранка–Николсона ($\theta = 1/2$) порядок аппроксимации схемы (6.109) составляет $O\left(\tau^2 + h^2\right)$, т. е. на один порядок по времени выше, чем для обычных явных или неявных схем. Таким образом, схема Кранка–Николсона (6.109) при $\theta = 1/2$ *абсолютно устойчива и имеет второй порядок аппроксимации по времени и пространственной переменной* x.

Аналогично схема Кранка–Николсона записывается и исследуется для других нестационарных уравнений, например для волнового уравнения.

6.6.3. Метод прямых. Метод прямых представляет собой метод частичной дискретизации, когда часть дифференциальных операторов, входящих в дифференциальную задачу, аппроксимируется с помощью отношения конечных разностей, а остальные дифференциальные операторы сохраняются в дифференциальной форме. Ниже рассматриваются два варианта метода прямых: поперечный и продольный варианты применительно к эволюционным задачам математической физики (теплопроводности, малых колебаний и т. п.) с простейшими граничными условиями первого рода.

В поперечном варианте метода прямых дифференциальные операторы по времени аппроксимируются с помощью отношения конечных разностей, дифференциальные операторы по пространственным переменным сохраняются, а решение отыскивается вдоль прямых $t = \text{const}$ (рис. 6.8).

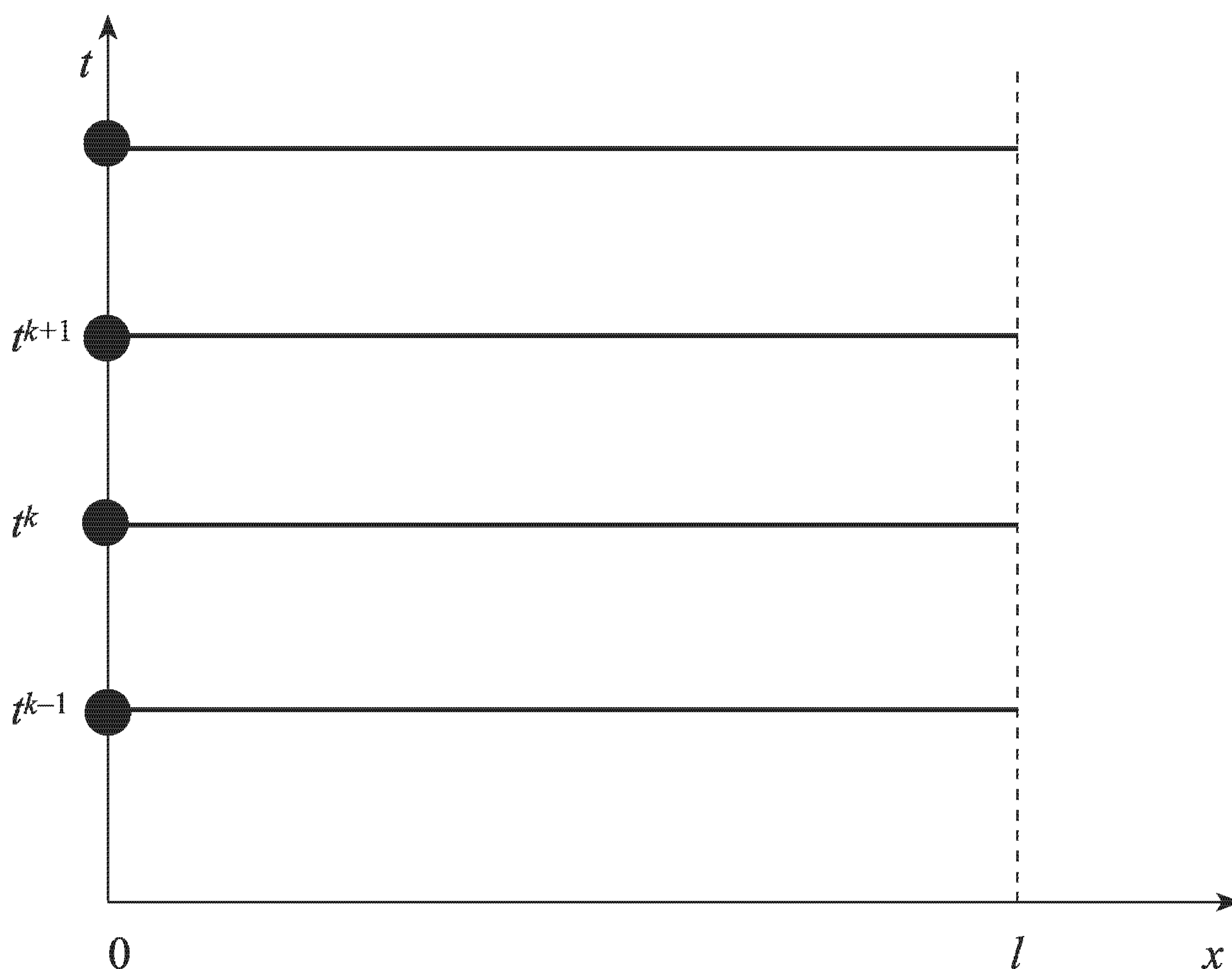


Рис. 6.8. Схема поперечного варианта метода прямых

В случае первой начально-краевой задачи для одномерного уравнения теплопроводности (6.1)–(6.4) этот вариант метода прямых имеет вид (в уравнении (6.1) коэффициент переноса

примем равным a)

$$a\tau \frac{d^2 u^{k+1}}{dx^2} - u^{k+1} = -u^k + O(\tau), \quad k = 0, 1, 2, \ldots, K, \tag{6.112}$$

$$u(0, t^{k+1}) = u^{k+1}(0) = \varphi_1(t^{k+1}), \quad k = 0, 1, 2, \ldots, K, \tag{6.113}$$

$$u(l, t^{k+1}) = u^{k+1}(l) = \varphi_2(t^{k+1}), \quad k = 0, 1, 2, \ldots, K, \tag{6.114}$$

т. е. дифференциальная задача (6.1)–(6.4) сведена к системе независимых краевых задач (6.112)–(6.114) для обыкновенных дифференциальных уравнений (ОДУ) 2-го порядка в количестве, равном числу $K + 1$, с граничными условиями первого рода.

Решением этой системы ОДУ, каждое из которых является ОДУ второго порядка, будут следующие явные функции:

$$u^{k+1}(x) = C_1 \exp(x/\sqrt{a\tau}\,) + C_2 \exp(-x/\sqrt{a\tau}\,) -$$
$$- 1/(2\sqrt{a\tau}\,) \exp(x/\sqrt{a\tau}\,) \cdot \int u^k(x) \exp(-x/\sqrt{a\tau}\,) dx +$$
$$+ 1/(2\sqrt{a\tau}\,) \exp(-x/\sqrt{a\tau}\,) \cdot \int u^k(x) \exp(x/\sqrt{a\tau}\,) dx,$$
$$k = 0, 1, 2, \ldots, K; \tag{6.115}$$

$$u^0(x) = u(x, 0) = \psi(x), \tag{6.116}$$

где постоянные интегрирования C_1, C_2 определяются путем подстановки (6.115) в краевые условия (6.113), (6.114) соответственно при $x = 0$ и $x = l$.

Аналогично, в случае первой начально-краевой задачи для волнового уравнения (6.12)–(6.16) схема поперечного варианта метода прямых на неявном шаблоне имеет вид

$$a^2\tau^2 \frac{\partial^2 u^{k+1}}{\partial x^2} - u^{k+1} = -2u^k + u^{k-1} + O(\tau),$$
$$k = 1, 2, \ldots, K; \tag{6.117}$$

$$u(x,0) = u^0 = \psi_1(x); \tag{6.118}$$

$$u(x,\tau) = u^1 = \psi_1(x) + \tau\psi_2(x) + O(\tau^2); \tag{6.119}$$

$$u(0,t^{k+1}) = u^{k+1}(0) = \varphi_1(t^{k+1}); \tag{6.120}$$

$$u(l,t^{k+1}) = u^{k+1}(l) = \varphi_2(t^{k+1}). \tag{6.121}$$

Аналитическое решение краевых задач для ОДУ 2-го порядка (6.117) также не вызывает затруднений.

В продольном варианте метода прямых с помощью отношения конечных разностей аппроксимируются пространственные дифференциальные операторы, а дифференциальные операторы по времени сохраняются. Решения отыскиваются вдоль прямых $x = \text{const}$ (рис. 6.9).

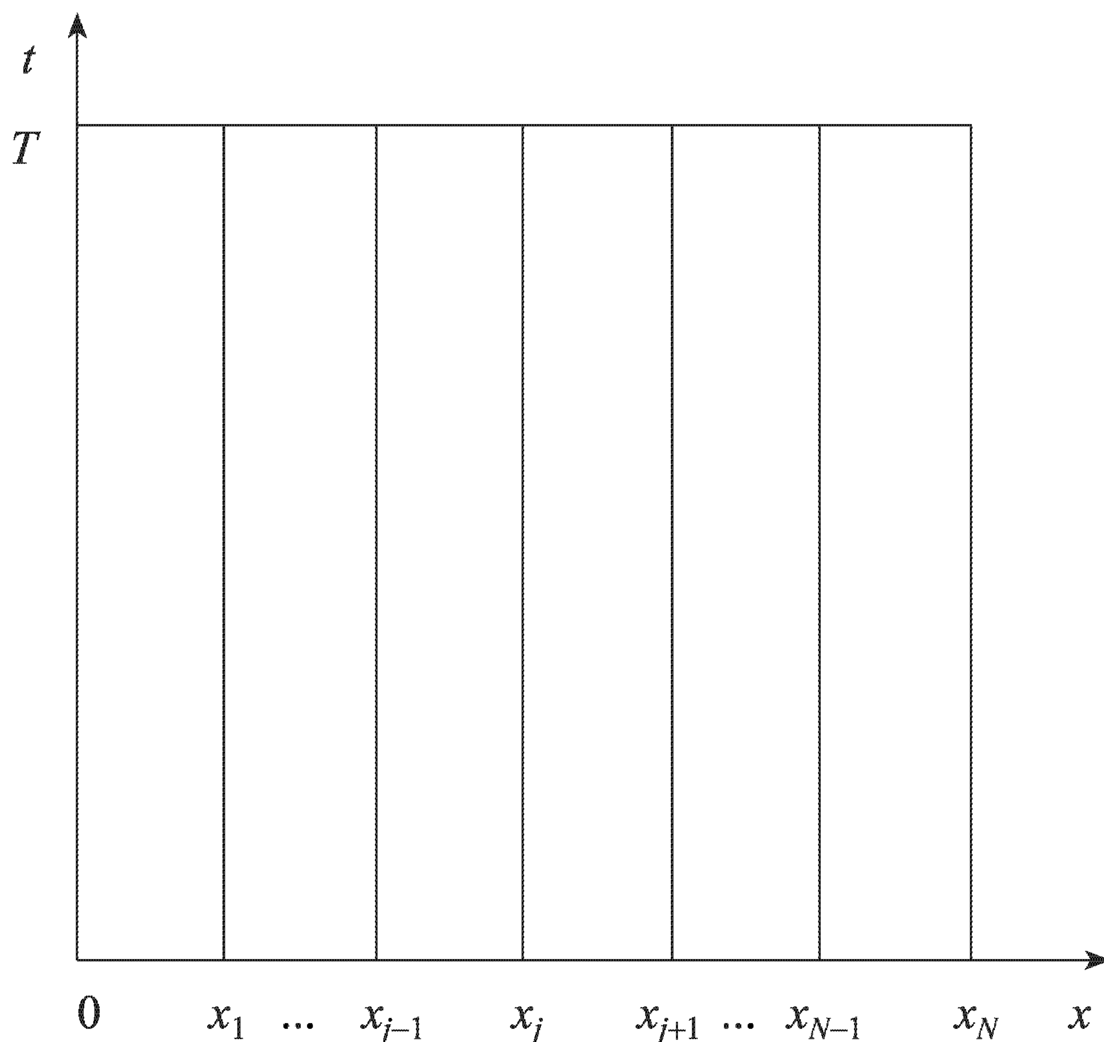


Рис. 6.9. Схема продольного варианта метода прямых

В результате получаем следующую задачу Коши для нормальной неоднородной системы ОДУ $(N-1)$-го порядка (на границах $x = 0$ и $x = l$ решения известны из левого (6.2) и правого (6.3) краевых условий соответственно):

$$
\left\{
\begin{aligned}
&\frac{du_1}{dt} = \frac{a^2}{h^2}\left(\varphi_1(t) - 2u_1 + u_2\right) + O(h^2),\\
&\frac{du_2}{dt} = \frac{a^2}{h^2}(u_1 - 2u_2 + u_3) + O(h^2),\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
&\frac{du_{N-1}}{dt} = \frac{a^2}{h^2}(u_{N-2} - 2u_{N-1} + \varphi_2(t)) + O\left(h^2\right),\\
&u_1(0) = \psi(x_1),\\
&u_2(0) = \psi(x_2),\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
&u_{N-1}(0) = \psi(x_{N-1}).
\end{aligned}
\right. \tag{6.122}
$$

Здесь искомыми являются функции $u_1(t),\ u_2(t), \dots, u_{N-1}(t)$. Задачу Коши (6.122) можно представить в следующей векторно-матричной форме:

$$
\begin{gathered}
u'(t) = Au + F(t), \quad u = (u_1\, u_2\, \dots\, u_{N-1})^T,\\
F(t) = \left(\frac{a^2}{h^2}\varphi_1(t)\ 0 \dots 0\ \frac{a^2}{h^2}\varphi_2\,(t)\right)^T,\\
u(0) = \psi(x), \quad x = (x_1\, x_2\, \dots\, x_{N-1})^T,
\end{gathered} \tag{6.123}
$$

$$
A = \begin{pmatrix}
-2a^2/h^2 & a^2/h^2 & & & \\
a^2/h^2 & -2a^2/h^2 & a^2/h^2 & & \\
\dots & \dots & \dots & \dots & \dots \\
 & & a^2/h^2 & -2a^2/h^2 & a^2/h^2 \\
 & & & a^2/h^2 & -2a^2/h^2
\end{pmatrix}.
$$

Ее решение можно найти методом вариации произвольных постоянных, для чего, интегрируя формально однородную си-

стему, соответствующую (6.123), получим общее решение однородной системы $u_{oo}(t)$ в виде матричной экспоненты

$$u_{\text{oo}}(t) = [\exp(At)]\, C, \quad C = (C_1 \ldots C_{N-1})^T\,,$$

в котором, разлагая матричную экспоненту в матричный ряд, получим

$$u_{\text{oo}}(t) = \left(E + At + \frac{1}{2}A^2t^2 + \frac{1}{6}A^3t^3 + \ldots\right) \cdot C = B(t) \cdot C,$$

где $B(t)$ — функциональная матрица, полученная суммированием ограниченного числа членов ряда.

Поскольку матрица $B(t)$ невырождена (матрица A невырождена), то определитель этой матрицы, являющийся определителем Вронского $W_B(t)$ фундаментальной системы решений однородной системы, соответствующей системе (6.123), не равен тождественно нулю, и, следовательно, в соответствии с методом вариации произвольных постоянных общее решение $u_{\text{он}}(t)$ неоднородной системы (6.123) выписывается в виде

$$u_{\text{он}}(t) = B(t)\widetilde{C}(t),$$

где $\widetilde{C}(t)$ — вектор-функция $\left(\widetilde{C}_1(t)\, \widetilde{C}_2(t)\, \ldots\, \widetilde{C}_{N-1}(t)\right)^T$, определяемая из решения СЛАУ

$$B(t) \cdot \widetilde{C}'(t) = F(t),$$

откуда

$$\widetilde{C}(t) = \int B^{-1}(t) \cdot F(t)dt + D,$$

где $D = (D_1\, D_2\, \ldots\, D_{N-1})^T$ — вектор произвольных постоянных.

Таким образом, общим решением системы (6.123) будет вектор

$$u(t) = B(t) \cdot \widetilde{C}(t),$$

$$\widetilde{C}(t) = \left(\int \frac{W_{B_1}(t)}{W_B(t)}dt + D_1 \qquad \int \frac{W_{B_2}(t)}{W_B(t)} + D_2 \qquad \ldots \qquad \ldots \qquad \int \frac{W_{B_{N-1}}(t)}{W_B(t)}dt + D_{N-1}\right)^T, \quad (6.124)$$

где $W_{Bj}(t)$, $j = \overline{1, N-1}$, — определители, получающиеся из определителя Вронского $W_B(t)$ заменой j-го столбца на вектор правых частей $F(t)$.

Аналогично можно применить продольный вариант метода прямых к решению первой начально-краевой задачи для волнового уравнения и к задаче Дирихле для уравнения Лапласа или Пуассона.

§ 6.7. Метод конечных разностей решения задач для волнового уравнения с граничными условиями, содержащими производные

Рассмотрим 3-ю начально-краевую задачу для волнового уравнения в общем виде:

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} + b\frac{\partial u}{\partial x} + cu, \quad 0 < x < l, \quad t > 0; \tag{6.125}$$

$$\frac{\partial u(0,t)}{\partial x} + \alpha_0 u(0,t) = \varphi_1(t), \quad x = 0, \quad t > 0; \tag{6.126}$$

$$\frac{\partial u(l,t)}{\partial x} + \alpha_1 u(l,t) = \varphi_2(t), \quad x = l, \quad t > 0; \tag{6.127}$$

$$u(x,0) = \psi_1(x), \quad 0 \leqslant x \leqslant l, \quad t = 0; \tag{6.128}$$

$$\frac{\partial u(x,0)}{\partial t} = \psi_2(x), \quad 0 \leqslant x \leqslant l, \quad t = 0. \tag{6.129}$$

Конечно-разностная схема для уравнения (6.125) на неявном шаблоне (рис. 6.3 *б*):

$$u_j^{k+1} - 2u_j^k + u_j^{k-1} = \sigma \left(u_{j+1}^{k+1} - 2u_j^{k+1} + u_{j-1}^{k+1} \right) +$$

$$+ \beta \left(u_{j+1}^{k+1} - u_{j-1}^{k+1} \right) + \gamma\, u_j^{k+1} + O(\tau + h^2), \quad j = \overline{1,\, N-1},$$

$$k = 1, 2, \ldots, \quad \sigma = a^2\tau^2/h^2, \quad \beta = b\tau^2/2h, \quad \gamma = c\tau^2, \tag{6.130}$$

имеет только первый порядок аппроксимации по времени, а на явном шаблоне (рис. 6.3 *а*):

$$u_j^{k+1} - 2u_j^k + u_j^{k-1} = \sigma\left(u_{j+1}^k - 2u_j^k + u_{j-1}^k\right) + \\ + \beta\left(u_{j+1}^k - u_{j-1}^k\right) + \gamma\, u_j^k + O(\tau^2 + h^2), \quad j = \overline{1,\, N-1}, \\ k = 1, 2, \ldots, \quad \sigma = a^2\tau^2/h^2, \quad \beta = b\tau^2/2h, \quad \gamma = c\tau^2,$$

— второй порядок. Второй порядок по времени имеет и неявно-явная схема с весом $\theta = 1/2$ (схема Кранка–Николсона):

$$\frac{u_j^{k+1} + 2u_j^k + u_j^{k-1}}{\tau^2} = \\ = \frac{1}{2}\left[a^2\frac{u_{j+1}^{k+1} - 2u_j^{k+1} + u_{j-1}^{k+1}}{h^2} + b\frac{u_{j+1}^{k+1} - u_{j-1}^{k+1}}{2h} + cu_j^{k+1}\right] + \\ + \frac{1}{2}\left[a^2\frac{u_{j+1}^{k-1} - 2u_j^{k-1} + u_{j-1}^{k-1}}{h^2} + b\frac{u_{j+1}^{k-1} - u_{j-1}^{k-1}}{2h} + cu_j^{k-1}\right] + \\ + O(\tau^2 + h^2), \quad j = \overline{1,\, N-1}, \quad k = 1, 2, \ldots$$

с шаблоном, представленным на рис. 6.10.

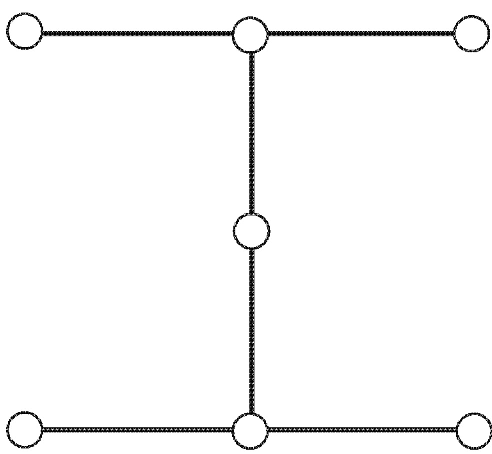


Рис. 6.10. Шаблон схемы Кранка–Николсона для волнового уравнения

При аппроксимации граничных условий (6.126), (6.127), содержащих производные по пространственным переменным, аппроксимация их с помощью отношения односторонних конечных разностей приводит не только к первому порядку аппроксимации по пространственной переменной (в этом случае аппроксимация всей начально-краевой задачи имеет порядок $O(h)$, несмотря на второй порядок аппроксимации во всех внутренних узлах расчетной области), но и к неконсервативности конечно-разностной схемы, поскольку на пространственных шагах, примыкающих к границам, не учитывается дифференциальное уравнение, т. е. на границах в конечно-разностной схеме не соблюдаются законы сохранения, на основе

которых выведено дифференциальное уравнение. Для устранения этого явления в конечно-разностной аппроксимации краевых условий (6.126), (6.127) учтем дифференциальное уравнение (6.125), для чего сделаем предположение о существовании на границах производной второго порядка искомой функции $u(x, t)$ по пространственной переменной, а затем разложим на точном решении в ряд Тейлора до третьего слагаемого включительно в окрестности граничных узлов значения сеточной функции u_1^{k+1} и u_{N-1}^{k+1} (т. е. в узлах, непосредственно примыкающих к граничным), получим

$$u_1^{k+1} = u_0^{k+1} + \left.\frac{\partial u}{\partial x}\right|_0^{k+1} h + \left.\frac{\partial^2 u}{\partial x^2}\right|_0^{k+1} \frac{h^2}{2} + O(h^3), \tag{6.131}$$

$$u_{N-1}^{k+1} = u_N^{k+1} - \left.\frac{\partial u}{\partial x}\right|_N^{k+1} h + \left.\frac{\partial^2 u}{\partial x^2}\right|_N^{k+1} \frac{h^2}{2} + O(h^3), \tag{6.132}$$

причем производные второго порядка в (6.131), (6.132) определим из дифференциального уравнения (6.125). В результате значения производных первого порядка $\left.\frac{\partial u}{\partial x}\right|_0^{k+1}$ и $\left.\frac{\partial u}{\partial x}\right|_N^{k+1}$ на границах определяются со вторым порядком следующим образом:

$$\left.\frac{\partial u}{\partial x}\right|_0^{k+1} = \frac{u_1^{k+1} - u_0^{k+1}}{h \cdot p} - \frac{h}{2a^2 \cdot p} \left.\frac{\partial^2 u}{\partial t^2}\right|_0^{k+1} + \frac{ch}{2a^2 p} u_0^{k+1} + \\ + O(h^2), \quad p = 1 - \frac{bh}{2a^2};$$

$$\left.\frac{\partial u}{\partial x}\right|_N^{k+1} = \frac{u_N^{k+1} - u_{N-1}^{k+1}}{h \cdot q} + \frac{h}{2a^2 \cdot q} \left.\frac{\partial^2 u}{\partial t^2}\right|_N^{k+1} - \frac{ch}{2a^2 q} u_N^{k+1} + \\ + O(h^2), \quad q = 1 + \frac{bh}{2a^2};$$

Подставляя их в граничные условия (6.126), (6.127), а затем аппроксимируя вторые производные по времени на верхнем $(k + 1)$-м временном слое (т. е. с порядком $O(\tau)$):

$$\left.\frac{\partial^2 u}{\partial t^2}\right|_0^{k+1} = \frac{1}{\tau^2}(u_0^{k+1} - 2u_0^k + u_0^{k-1}) + O(\tau),$$

$$\left.\frac{\partial^2 u}{\partial t^2}\right|_N^{k+1} = \frac{1}{\tau^2}(u_N^{k+1} - 2u_N^k + u_N^{k-1}) + O(\tau),$$

получим алгебраические уравнения для граничных узлов, каждое из которых содержит два неизвестных, в следующей форме:

$$b_0 u_0^{k+1} + c_0 u_1^{k+1} = d_0 + O(\tau + h^2); \qquad (6.133)$$

$$a_N u_{N-1}^{k+1} + b_N u_N^{k+1} = d_N + O(\tau + h^2), \qquad (6.134)$$

где

$$a_0 = 0;$$

$$b_0 = -\frac{1}{hp} - \frac{h}{2a^2 p\tau^2} + \frac{ch}{2a^2 p} + \alpha_0;$$

$$c_0 = \frac{1}{hp};$$

$$d_0 = \varphi_1(t^{k+1}) + \frac{h}{2a^2 p\tau^2}\left(-2u_0^k + u_0^{k-1}\right);$$

$$b_N = \frac{1}{hq} + \frac{h}{2a^2 q\tau^2} - \frac{ch}{2a^2 q} + \alpha_1;$$

$$a_N = -\frac{1}{hq};$$

$$d_N = \varphi_2(t^{k+1}) - \frac{h}{2a^2 q\tau^2}\left(-2u_N^k + u_N^{k-1}\right);$$

$$c_N = 0.$$

Для внутренних узлов расчетной сетки неявная конечно-разностная схема (6.130) представляется в виде

$$a_j u_{j-1}^{k+1} + b_j u_j^{k+1} + c_j u_{j+1}^{k+1} = d_j + O(\tau + h^2),$$

$$j = \overline{1, N-1}, \quad k = 1, 2, \ldots, \qquad (6.135)$$

где

$$a_j = \sigma - \beta;$$

$$b_j = -2\sigma - 1 + \gamma;$$

$$c_j = \sigma + \beta;$$

$$d_j = -2u_j^k + u_j^{k-1}.$$

СЛАУ (6.133)–(6.135) имеет трехдиагональную матрицу, что позволяет решать ее на каждом временном слое методом прогонки. При $k = 1$ u_j^0 и u_j^1, $j = \overline{0, N}$, вычисляются в соответствии с начальными условиями (6.128), (6.129):

$$u_j^0 = \psi_1(x_j), \quad j = \overline{0, N};$$

$$u_j^1 = u_j^0 + \tau\psi_2(x_j) + O(\tau^2) =$$
$$= \psi_1(x_j) + \tau\psi_2(x_j) + O(\tau^2), \quad j = \overline{0, N}.$$

Аналогично строятся конечно-разностные аппроксимации со вторым порядком по пространственной переменной на явном (рис. 6.3 *а*) и неявно-явном шаблонах, в частности на шаблоне (рис. 6.10) для схемы Кранка–Николсона.

§ 6.8. Метод установления и его обоснование

Из теоретических основ механики сплошных сред, и, в частности, из газовой динамики, известно, что в *сверхзвуковых* течениях, описываемых уравнениями *гиперболического типа* возмущения газодинамических функций, возникающие в какой-либо точке (например, погрешности численного решения), локализуются в так называемом сверхзвуковом конусе Маха, расположенном вниз по течению от точки возмущения, с углом раствора, обратно пропорциональным числу Маха (число Маха — отношение скорости течения к скорости звука в среде). В *околозвуковых* (*трансзвуковых*) течениях, описываемых уравнениями *параболического типа*, возмущения локализуются в окрестности точки возмущения. По этой причине большинство встречающихся на практике параболических задач хорошо обусловлены по устойчивости и сходимости.

В *дозвуковых газодинамических течениях*, описываемых уравнениями *эллиптического типа*, возмущения, возникающие в точке, передаются во все стороны. В соответствии с этим численные методы в задачах для уравнений эллиптического типа

плохо обусловлены, поскольку погрешности счета с одинаковой скоростью передаются как в область, где уже проведен расчет, так и в область, где расчет еще не проводился. Это затрудняет применение к эллиптическим задачам так называемых маршевых методов, когда численное решение осуществляется последовательным переходом от одного пространственного сечения к другому.

Этот факт стимулировал поиск путей перехода от эллиптических задач к *эквивалентным* (при определенных условиях) параболическим задачам. Оказалось, что решение параболических задач при $t \to \infty$ асимптотически стремится к решению соответствующих эллиптических задач, т. е. к решению нестационарных задач, в которых искомая функция уже не зависит от времени.

В соответствии с этой особенностью нестационарных задач был разработан и обоснован *метод установления*. В его разработке и обосновании активное участие принимали отечественные математики, и, в частности, С. К. Годунов.

Рассмотрим обоснование метода установления на примере задачи Дирихле для уравнения Пуассона:

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = f(x, y), \quad (x, y) \in \Omega, & (6.136) \\[2ex] u(x, y)\big|_{\Gamma} = \varphi(x, y), \quad (x, y) \in \Gamma. & (6.137) \end{cases}$$

В соответствии с методом установления вместо задачи (6.136), (6.137) рассмотрим следующую начально-краевую задачу для уравнения параболического типа (двумерного уравнения теплопроводности), добавив в правую часть уравнения (6.136) дифференциальный оператор $\dfrac{\partial u}{\partial t}$ и дополняя входные данные однородным начальным условием, сохранив при этом независимость от времени краевого условия (6.137):

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - f(x, y), \qquad (x, y) \in \Omega, \tag{6.138}$$

$$u(x,y,t)\big|_{\Gamma} = \varphi(x,y),\qquad (x,y)\in\Gamma, \tag{6.139}$$

$$u(x,y,0)\big|_{\overline{\Omega}} = 0,\qquad (x,y)\in\overline{\Omega} = \Omega+\Gamma, \tag{6.140}$$

где Ω — плоская двумерная область, ограниченная границей Γ.

Покажем, что при $t\to\infty$ решение задачи (6.138)–(6.140) стремится к решению задачи (6.136), (6.137). С этой целью для задачи (6.136), (6.137) введем пространственную сетку ω_h, на которой дифференциальную задачу (6.136), (6.137) будем аппроксимировать с помощью отношения конечных разностей, в результате чего получим следующее векторно-матричное уравнение:

$$A\alpha = f, \tag{6.141}$$

т. е. СЛАУ для функции $\alpha(x_i,y_j)$.

В соответствии с продольным вариантом метода прямых для задачи (6.138)–(6.140) можно записать следующую задачу Коши для векторно-матричного ОДУ 1-го порядка:

$$\begin{cases} \dfrac{d\psi}{dt} + A\psi = f, & (6.142)\\[2ex] \psi(0) = 0, & (6.143)\end{cases}$$

для функции $\psi(x_i,y_j,t)$.

Параллельно рассмотрим задачу на собственные значения и собственные векторы матрицы A:

$$Au^k = \lambda_k u^k,\quad k=\overline{1,n}, \tag{6.144}$$

где u^k — собственные векторы, λ_k — собственные значения матрицы A.

Если матрица A — симметрическая $\left(A = A^{\mathrm{T}}\right)$, положительно определенная $((Au,u)>0)$, а именно к таким матрицам чаще всего приводит конечно-разностная аппроксимация дифференциальных задач, то имеется полный вещественный положительный спектр $\lambda_k>0$, $k=\overline{1,n}$, а собственные векторы u^k, $k=\overline{1,n}$, ортогональны.

Разложим функции ψ, α, f в ряд по собственным векторам u^k, получим

$$\alpha = \sum_{k=1}^{n} \alpha_k \cdot u^k, \tag{6.145}$$

$$\psi = \sum_{k=1}^{n} \psi_k \cdot u^k, \tag{6.146}$$

$$f = \sum_{k=1}^{n} f_k \cdot u^k, \tag{6.147}$$

где $\alpha_k = (\alpha, u^k)$, $\psi_k = (\psi, u^k)$, $f_k = (f, u^k)$ — коэффициенты разложений (6.145)–(6.147).

Для определения коэффициентов α_k разложения решения конечно-разностной задачи (6.141) подставим (6.145) и (6.147) в (6.141), получим

$$A\left(\sum_{k=1}^{n} \alpha_k \cdot u^k\right) = \sum_{k=1}^{n} f_k \cdot u^k,$$

а поскольку

$$A\left(\sum_{k=1}^{n} \alpha_k \cdot u^k\right) = \sum_{k=1}^{n} A(\alpha_k \cdot u^k) =$$
$$= \sum_{k=1}^{n} \alpha_k (A u^k) = \sum_{k=1}^{n} \alpha_k \cdot (\lambda_k u^k),$$

то

$$\sum_{k=1}^{n} (\alpha_k \lambda_k) \cdot u^k = \sum_{k=1}^{n} f_k u^k,$$

откуда следует, что

$$\alpha_k = \frac{f_k}{\lambda_k}, \quad k = \overline{1, n}, \tag{6.148}$$

поскольку u^k, $k = \overline{1, n}$, ортогональны, и, следовательно, *линейно независимы*.

Таким образом, решение (6.145) стационарной задачи (6.136), (6.137) с учетом (6.148) приобретает вид

$$\alpha = \sum_{k=1}^{n} \frac{f_k}{\lambda_k} u^k. \qquad (6.149)$$

Подставим теперь (6.146), (6.147) в задачу Коши (6.142), (6.143), получим

$$\frac{d}{dt}\left(\sum_{k=1}^{n} \psi_k u^k\right) + A\left(\sum_{k=1}^{n} \psi_k u^k\right) = \sum_{k=1}^{n} f_k u^k,$$

$$\sum_{k=1}^{n} \psi_k(0) \cdot u^k = 0.$$

Поскольку собственные векторы u^k линейно независимы, то, учитывая равенство $Au^k = \lambda_k u^k$, получим следующую задачу Коши для функции $\psi_k(t)$:

$$\frac{d\psi_k}{dt} + \lambda_k \psi_k = f_k, \quad k = \overline{1, n}, \qquad (6.150)$$

$$\psi_k(0) = 0, \quad k = \overline{1, n}, \qquad (6.151)$$

решением которой является функция

$$\psi_k(t) = \frac{f_k}{\lambda_k} \left(1 - \exp\left(-\lambda_k t\right)\right). \qquad (6.152)$$

Подставляя (6.152) в разложение (6.146), получим решение задачи (6.142), (6.143), а следовательно, и задачи (6.138)–(6.140) в виде

$$\psi(t) = \sum_{k=1}^{n} \frac{f_k}{\lambda_k} (1 - \exp(-\lambda_k t)) \cdot u^k. \qquad (6.153)$$

Для положительно определенных матриц A все собственные значения положительны, и, следовательно, при $t \to \infty$ из (6.153) получаем решение

$$\psi = \sum_{k=1}^{n} \frac{f_k}{\lambda_k} \cdot u^k, \qquad (6.154)$$

совпадающее с решением (6.149) стационарной задачи (6.136), (6.137).

Таким образом, вместо решения задачи Дирихле (6.136), (6.137) можно решать нестационарную задачу (6.138)–(6.140) для двумерного уравнения теплопроводности при больших значениях времени. Решение останавливается тогда, когда оно уже несущественно изменяется по времени.

При этом для нестационарной задачи (6.138)–(6.140) можно использовать богатейший арсенал численных методов, включая методы расщепления.

ГЛАВА VII

МЕТОД КОНЕЧНЫХ РАЗНОСТЕЙ РЕШЕНИЯ МНОГОМЕРНЫХ ЗАДАЧ МАТЕМАТИЧЕСКОЙ ФИЗИКИ. МЕТОДЫ РАСЩЕПЛЕНИЯ

Программа

Экономичность конечно-разностных схем для многомерных уравнений математической физики. Методы матричной прогонки, переменных направлений Писмена–Рэчфорда, дробных шагов Н. Н. Яненко, центрально-симметричный А. А. Самарского, переменных направлений с экстраполяцией В. Ф. Формалева и полного расщепления Формалева–Тюкина. Методы расщепления численного решения эллиптических задач. Численные методы решения задач для многомерных уравнений гиперболического типа: метод характеристик решения квазилинейных гиперболических систем, метод С. К. Годунова. Задача о распаде произвольного разрыва.

При численном решении многомерных задач математической физики исключительно важным является вопрос об *экономичности* используемых методов.

Конечно-разностную схему будем называть экономичной, если число длинных операций (операций типа умножения) пропорционально числу узлов сетки.

За последние 50 лет разработано значительное количество экономичных разностных схем численного решения многомерных задач математической физики, основанных на *расщеплении* пространственных дифференциальных операторов по координатным направлениям и использовании метода скалярной прогонки вдоль этих направлений. При этом большинство разработанных схем активно используют явную аппроксимацию дифференциальных операторов, и поэтому при определенных соотношениях сеточных характеристик в задачах с граничными условиями, содержащими производные, или в задачах, содержащих смешанные производные, такие схемы становятся условно устойчивыми [10–16].

Из экономичных конечно-разностных схем, получивших наибольшее распространение, в данной главе рассматриваются следующие:

— схема *метода переменных направлений* Писмена–Рэчфорда (1955 г.);

— схема *метода дробных шагов* Н. Н. Яненко (1956 г.);

— схема *центрально-симметричного метода* А. А. Самарского (1958 г.).

Эти схемы аппроксимируют смешанные дифференциальные операторы на нижних временных слоях (явным образом), что при определенных условиях может приводить к неустойчивости решения. От этих недостатков свободны следующие экономичные, абсолютно устойчивые методы:

— *метод переменных направлений с экстраполяцией* В. Ф. Формалева (1987 г.) [17–19];

— *метод полного расщепления* Формалева–Тюкина (1996 г.) [20–21].

Все эти методы будем называть общим термином — *методы расщепления*.

Среди неэкономичных, полностью неявных, а потому абсолютно устойчивых, методов можно отметить *метод матричной прогонки*.

Рассмотрим эти методы на примере задачи для двумерного уравнения параболического типа в прямоугольнике со сторонами ℓ_1, ℓ_2 и граничными условиями 1-го рода.

Рассматриваемые методы можно использовать и для эллиптических задач, если с помощью метода установления (§ 6.8) свести их к параболическим задачам, а также для гиперболических задач.

Для пространственно-временной области $\overline{\mathrm{G}}_\mathrm{T} = \overline{\mathbf{G}} \times [0, \mathrm{T}]$, $t \in [0, \mathrm{T}]$, $\overline{\mathrm{G}} = \mathrm{G} + \Gamma$, $\mathrm{G} = \ell_1 \times \ell_2$, рассмотрим следующую задачу:

$$\frac{\partial u}{\partial t} = a\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \quad x \in (0, \ell_1), \quad y \in (0, \ell_2), \quad t > 0; \qquad (7.1)$$

$$u(x, 0, t) = \varphi_1(x, t), \quad x \in [0,\ \ell_1], \quad y = 0, t > 0; \qquad (7.2)$$

$$u(x, \ell_2, t) = \varphi_2(x, t), \quad x \in [0,\ \ell_1], \quad y = \ell_2, \quad t > 0; \qquad (7.3)$$

$$u(0, y, t) = \varphi_3(y, t), \quad x = 0,\ y \in [0, \ell_2], \quad t > 0; \qquad (7.4)$$

$$u(\ell_1, y, t) = \varphi_4(y, t), \quad x = \ell_1,\ y \in [0, \ell_2], \quad t > 0; \tag{7.5}$$

$$u(x, y, 0) = \psi(x, y), \quad x \in [0, \ell_1], \quad y \in [0, \ell_2], \quad t = 0. \tag{7.6}$$

Введем следующую пространственно-временную сетку с шагами h_1, h_2, τ соответственно по переменным x, y, t:

$$\omega^{\tau}_{h_1, h_2} = \Big\{ x_i = ih_1, i = \overline{0, I};\ x_j = jh_2,$$

$$j = \overline{0, J};\ \ t^k = k\tau,\ \ k = 0, 1, 2, \ldots \Big\}, \tag{7.7}$$

и на этой сетке будем аппроксимировать дифференциальную задачу (7.1)–(7.6) методом конечных разностей на верхнем временном слое $t^{k+1} = (k+1)\tau$:

$$\frac{u_{ij}^{k+1} - u_{ij}^{k}}{\tau} = \frac{a}{h_1^2}\left(u_{i+1j}^{k+1} - 2u_{ij}^{k+1} + u_{i-1j}^{k+1}\right) +$$

$$+ \frac{a}{h_2^2}\left(u_{ij+1}^{k+1} - 2u_{ij}^{k+1} + u_{ij-1}^{k+1}\right),$$

$$i = \overline{1, I-1}\ ;\ j = \overline{1, J-1};\ \ k = 0, 1, 2, \ldots\ ; \tag{7.8}$$

$$u_{i\,0}^{k+1} = \varphi_1(x_i, t^{k+1});\ u_{i\,J}^{k+1} = \varphi_2(x_i, t^{k+1});\ u_{0\,j}^{k+1} = \varphi_3(y_j, t^{k+1});$$

$$u_{I\,j}^{k+1} = \varphi_4(y_j, t^{k+1});\ u_{ij}^{0} = \psi(x_i, y_j),\ i = \overline{0, I};\ \ j = \overline{0, J}.$$

§ 7.1. Метод матричной прогонки

Выберем три произвольных сечения $i-1, i, i+1, i = \overline{1, I-1}$, и запишем неявную схему (7.8) в следующей векторно-матричной форме:

$$-A_i u_{i-1}^{k+1} + B_i u_i^{k+1} - C_i u_{i+1}^{k+1} = F_i, \quad i = \overline{1, I-1}, \tag{7.9}$$

где A_i, B_i, C_i — квадратные матрицы размера $J-1$ вида

$$A_i = C_i = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_1 \end{bmatrix},$$

$$B_i = \begin{bmatrix} 1+2\sigma_1+ \\ +2\sigma_2 & -\sigma_2 & 0 & \ldots & 0 \\ -\sigma_2 & 1+2\sigma_1+ \\ +2\sigma_2 & -\sigma_2 & \ldots & 0 \\ 0 & -\sigma_2 & 1+2\sigma_1+ \\ +2\sigma_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1+2\sigma_1+ \\ +2\sigma_2 \end{bmatrix},$$

$\sigma_1 = \dfrac{a\tau}{h_1^2}$, $\sigma_2 = \dfrac{a\tau}{h_2^2}$ — сеточные числа Куранта; u_{i-1}^{k+1}, u_i^{k+1}, u_{i+1}^{k+1}, $i = \overline{1, I-1}$, — векторы значений сеточной функции размера $J-1$, $u_i^{k+1} = \left(u_{i\,1}^{k+1}\, u_{i\,2}^{k+1}\, \ldots\, u_{i\,J-1}^{k+1}\right)^T$; F_i, $i = \overline{1, I-1}$, — векторы правых частей размера $J-1$, $F_i = (f_{i\,1}\, f_{i2} \ldots f_{i\,J-1})^T$.

Из векторно-матричной СЛАУ (7.9) видно, что матрица СЛАУ представляет собой трехдиагональную матрицу, на главной диагонали которой стоят матрицы B_i, $i = \overline{1, I-1}$, на нижней диагонали — матрицы A_i, $i = \overline{1, I-1}$, а на верхней диагонали — матрицы C_i,

$$\begin{bmatrix} B_1 & C_1 & & & & & \\ A_2 & B_2 & C_2 & & & \theta & \\ & A_3 & B_3 & C_3 & & & \\ & & \ldots & \ldots & \ldots & & \\ & & & A_i & B_i & C_i & \\ & \theta & & & \ldots & \ldots & \ldots \\ & & & & & A_{I-1} & B_{I-1} \end{bmatrix} \begin{bmatrix} u_1^{k+1} \\ u_2^{k+1} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ u_{I-1}^{k+1} \end{bmatrix} = \begin{bmatrix} \widetilde{F}_1 \\ F_2 \\ F_3 \\ \vdots \\ \vdots \\ F_{I-2} \\ \widetilde{F}_{I-1} \end{bmatrix}, \qquad (7.10)$$

$A_1 = \theta$, $C_{I-1} = \theta$, θ — нулевая матрица.

Матрица A_1 равна нулевой матрице, поскольку при $i = 1$ первое матричное уравнение системы (7.9) имеет вид

$$B_1 u_1^{k+1} - C_1 u_2^{k+1} = \widetilde{F}_1, \quad \widetilde{F}_1 = F_1 + A_1 u_0^{k+1},$$

$$u_0^{k+1} = \varphi_3\left(y_j,\ t^{k+1}\right), \quad j = \overline{1, J-1}, \quad (7.11)$$

т. е. после вычисления вектора $\widetilde{F}_1$ с помощью матрицы A_1 в уравнении (7.11) матрицу A_1 можно положить равной нулевой матрице θ.

Аналогично, матрица $C_{\text{I - 1}}$ равна нулевой матрице, поскольку последнее уравнение системы (7.9) имеет вид

$$-A_{I-1}u_{I-2}^{k+1} + B_{I-1}u_{I-1}^{k+1} = \widetilde{F}_{I-1}, \quad \widetilde{F}_{I-1} = F_{I-1} + C_{I-1}u_I^{k+1},$$

$$u_I^{k+1} = \varphi_4\left(y_j, t^{k+1}\right), \quad j = \overline{1, J-1}, \quad (7.12)$$

т. е. после вычисления вектора $\widetilde{F}_{I-1}$ с помощью матрицы C_{I-1} последнюю можно принять равной нулевой матрице θ.

Как видно теперь, система (7.9) или (7.10) по форме совпадает с системой (2.1) для одного координатного направления, и эту систему формально можно решить с помощью метода прогонки. Будем искать решение системы (7.9) в виде

$$u_i^{k+1} = P_i\, u_{i+1}^{k+1} + Q_i, \quad i = \overline{1, I-1}, \quad (7.13)$$

где P_i — прогоночные матрицы, Q_i —прогоночные векторы.

Для определения P_i, Q_i выразим из (7.11) u_1^{k+1} через u_2^{k+1}, получим

$$u_1^{k+1} = B_1^{-1}C_1u_2^{k+1} + B_1^{-1}\widetilde{F}_1,$$

откуда, сравнивая с (7.13) при $i = 1$, находим

$$P_1 = B_1^{-1}C_1, \quad Q_1 = B_1^{-1}\widetilde{F}_1, \quad u_1^{k+1} = P_1\, u_2^{k+1} + Q_1. \quad (7.14)$$

Выразим из 2-го уравнения системы (7.9) u_2^{k+1} через u_3^{k+1}:

$$-A_2\left(P_1\, u_2^{k+1} + Q_1\right) + B_2u_2^{k+1} - C_2u_3^{k+1} = F_2,$$

откуда

$$u_2^{k+1} = \left(B_2 - A_2P_1\right)^{-1} C_2u_3^{k+1} + \left(B_2 - A_2P_1\right)^{-1}\left(F_2 + A_2Q_1\right),$$

$$P_2 = (B_2 - A_2P_1)^{-1}C_2; \; Q_2 = \left(B_2 - A_2P_1\right)^{-1}\left(F_2 + A_2Q_1\right). \quad (7.15)$$

Продолжая этот процесс, получим

$$P_i = \left(B_i - A_iP_{i-1}\right)^{-1} C_i; \; Q_i = \left(B_i - A_iP_{i-1}\right)^{-1}\left(F_i + A_iQ_{i-1}\right),$$

$$i = \overline{1, I-1}, \quad (7.16)$$

причем при $i = 1$ эти выражения трансформируются в (7.14), поскольку $A_1 = \theta$. При $i = I - 1$, $P_{I-1} = \theta$, поскольку $C_{I-1} = \theta$, тогда

$$Q_{I-1} = (B_{I-1} - A_{I-1}P_{I-2})^{-1}(F_{I-1} + A_{I-1}Q_{I-2}). \qquad (7.17)$$

После вычисления выражений (7.16), (7.17) прямой ход метода матричной прогонки завершен. Обратный ход по определению векторов u_i, $i = I - 1,\ I - 2,\ \ldots,\ 1$, осуществляется с помощью равенств (7.13):

$$\begin{cases} u_{I-1}^{k+1} = P_{I-1}u_I^{k+1} + Q_{I-1} = Q_{I-1}, \\ u_{I-2}^{k+1} = P_{I-2}u_{I-1}^{k+1} + Q_{I-2}, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ u_1^{k+1} = P_1 u_2^{k+1} + Q_1. \end{cases} \qquad (7.18)$$

К *достоинству* метода матричной прогонки относится его абсолютная устойчивость, поскольку аппроксимация осуществляется на верхнем временном слое (неявно) с порядком

$$O(\tau + |h|^2), \quad |h|^2 = h_1^2 + h_2^2,$$

а к *недостатку* — его неэкономичность, поскольку метод включает в себя обращение матриц вида $P_i = B_i - A_i P_{i-1}$.

§ 7.2. Метод переменных направлений Писмена–Рэчфорда

В схеме метода переменных направлений (МПН), как и во всех методах расщепления, шаг по времени τ разбивается на число, равное числу независимых пространственных переменных (в двумерном случае — на два). На каждом дробном временном слое один из пространственных дифференциальных операторов аппроксимируется неявно (по соответствующему координатному направлению осуществляются скалярные прогонки), а остальные явно. На следующем дробном шаге следующий по порядку дифференциальный оператор аппроксимируется неявно, а остальные — явно и т. д. В двумерном случае схема метода переменных направлений для задачи (7.1)–(7.6) имеет вид

$$\frac{u_{ij}^{k+1/2}-u_{ij}^{k}}{\tau/2}=\frac{a}{h_1^2}\left(u_{i+1\,j}^{k+1/2}-2u_{ij}^{k+1/2}+u_{i-1\,j}^{k+1/2}\right)+\frac{a}{h_2^2}\left(u_{ij+1}^{k}-\right.$$
$$\left.-2u_{ij}^{k}+u_{ij-1}^{k}\right), \quad (7.19)$$

$$\frac{u_{ij}^{k+1}-u_{ij}^{k+1/2}}{\tau/2}=\frac{a}{h_1^2}\left(u_{i+1\,j}^{k+1/2}-2u_{ij}^{k+1/2}+u_{i-1\,j}^{k+1/2}\right)+\frac{a}{h_2^2}\left(u_{ij+1}^{k+1}-\right.$$
$$\left.-2u_{ij}^{k+1}+u_{ij-1}^{k+1}\right). \quad (7.20)$$

В подсхеме (7.19) на первом дробном шаге $\tau/2$ оператор $a\dfrac{\partial^2}{\partial x^2}$ аппроксимируется неявно, а оператор $a\dfrac{\partial^2}{\partial y^2}$ — явно (в результате весь конечно-разностный оператор по переменной y переходит в правые части, поскольку u_{ij}^{k} известно). С помощью скалярных прогонок в количестве, равном числу $J-1$, в направлении переменной x получаем распределение сеточной функции $u_{ij}^{k+1/2}$, $i=\overline{1,I-1}$, $j=\overline{1,J-1}$, на временном полуслое $t^{k+1/2}=t^{k}+\tau/2$.

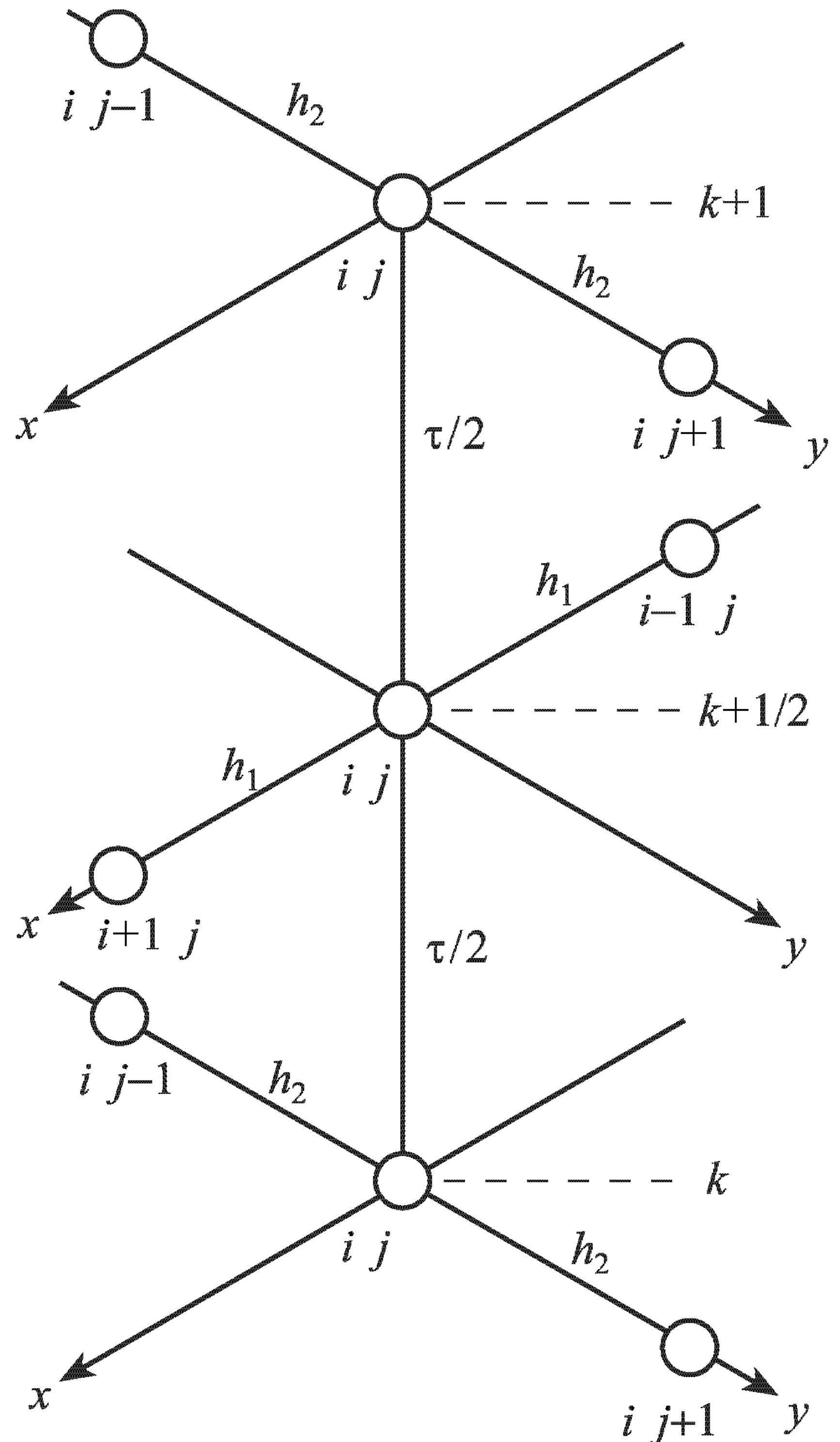


Рис. 7.1. Шаблон схемы метода переменных направлений

В подсхеме (7.20) оператор $a\dfrac{\partial^2}{\partial y^2}$ аппроксимируется неявно на верхнем временном слое $t^{k+1}=(k+1)\tau$, а оператор $a\dfrac{\partial^2}{\partial x^2}$ — явно в момент времени $t^{k+1/2}=t^{k}+\tau/2$ (конечно-разностный аналог этого оператора переходит в правые части). С помощью скалярных прогонок в направлении переменной y в количестве, равном числу $I-1$, получаем распределение сеточной функции u_{ij}^{k+1}, $i=\overline{1,I-1}$,

$j = \overline{1, J-1}$, на полуслое $t^{k+1} = t^{k+1/2} + \tau/2$. Шаблон схемы МПН представлен на рис. 7.1.

Исследуем схему МПН (7.19) , (7.20) на порядок аппроксимации и устойчивость, для чего запишем ее в следующей операторной форме:

$$\frac{u^{k+1/2} - u^k}{\tau/2} = \Lambda_1 u^{k+1/2} + \Lambda_2 u^k, \tag{7.21}$$

$$\frac{u^{k+1} - u^{k+1/2}}{\tau/2} = \Lambda_1 u^{k+1/2} + \Lambda_2 u^{k+1}, \tag{7.22}$$

где $\Lambda_1 = a\dfrac{\Delta^2}{h_1^2}$, $\Lambda_2 = a\dfrac{\Delta^2}{h_2^2}$, $u = (u_{11}u_{12}\ldots u_{I-1J-1})^T$.

Исключим из равенств (7.21), (7.22) сеточную функцию $u^{k+1/2}$ на промежуточном временном слое, для чего запишем их в виде

$$\left(E - \frac{\tau}{2}\Lambda_1\right) u^{k+1/2} = \left(E + \frac{\tau}{2}\Lambda_2\right) u^k, \tag{7.23}$$

$$\left(E - \frac{\tau}{2}\Lambda_2\right) u^{k+1} = \left(E + \frac{\tau}{2}\Lambda_1\right) u^{k+1/2}, \tag{7.24}$$

где E — тождественный оператор (например, единичная матрица). Умножим (7.23) слева на оператор $\left(E + \frac{\tau}{2}\Lambda_1\right)$, а (7.24) — на оператор $\left(E - \frac{\tau}{2}\Lambda_1\right)$ и, складывая результаты, получим

$$\left(E - \frac{\tau}{2}\Lambda_1\right)\left(E - \frac{\tau}{2}\Lambda_2\right) u^{k+1} = \left(E + \frac{\tau}{2}\Lambda_1\right)\left(E + \frac{\tau}{2}\Lambda_2\right) u^k, \tag{7.25}$$

т. е. (7.25) является уже двухслойной схемой. Раскроем в (7.25) скобки и разделим полученное выражение на τ:

$$\frac{u^{k+1} - u^k}{\tau} = \frac{1}{2}\left(\Lambda_1 u^{k+1} + \Lambda_1 u^k\right) + \frac{1}{2}\left(\Lambda_2 u^{k+1} + \Lambda_2 u^k\right) - \\ - \frac{\tau}{4}\Lambda_1\Lambda_2\left((u^{k+1} - u^k\right). \tag{7.26}$$

В векторно-операторном равенстве (7.26) первые два слагаемых в правой части аппроксимируют дифференциальные операторы $\dfrac{\partial^2 u}{\partial x^2}$ и $\dfrac{\partial^2 u}{\partial y^2}$ по схеме Кранка–Николсона с порядком $O(\tau^2 + |h|^2)$, а последнее слагаемое можно представить

как $\frac{\tau^2}{4}\Lambda_1\Lambda_2 u_t = O(\tau^2)$, поскольку $u_t = \frac{u^{k+1}-u^k}{\tau}$ — конечно-разностная производная по времени. Поэтому

$$\frac{u^{k+1}-u^k}{\tau} = \frac{1}{2}\left(\Lambda_1 u^{k+1} + \Lambda_1 u^k\right) + \frac{1}{2}\left(\Lambda_2 u^{k+1} + \Lambda_2 u^k\right) + \\ + O\left(\tau^2 + |h|^2\right),$$

т. е. схема МПН имеет второй порядок по времени и поэтому является высокоточной.

Исследуем устойчивость схемы (7.19) , (7.20) методом гармонического анализа, для чего вместо значений сеточной функции подставим гармонику

$$u_{ij}^k = \eta_k \exp\left[i\left(\lambda_n x_i + \lambda_m y_j\right)\right], \tag{7.27}$$

получим для подсхемы (7.19) равенство

$$\eta_{k+1/2} - \eta_k = \sigma_1 \eta_{k+1/2}\left[\exp(i\lambda_n h_1) - 2 + \exp\left(-i\lambda_n h_1\right)\right] + \\ + \sigma_2 \eta_k \left[\exp\left(i\lambda_m h_2\right) - 2 + \exp\left(-i\lambda_m h_2\right)\right],$$

откуда (см. п. 6.5.2)

$$\left|\frac{\eta_{k+1/2}}{\eta_k}\right| = \left|\frac{1-\sigma_2\, b_2}{1+\sigma_1 b_1}\right|,$$

где $b_1 = 4\,\sin^2\frac{\lambda_n h_1}{2}$; $b_2 = 4\,\sin^2\frac{\lambda_m h_2}{2}$; $\sigma_1 = \frac{a\tau}{h_1^2}$; $\sigma_2 = \frac{a\tau}{h_2^2}$.

Аналогично, подставляя гармонику (7.27) в подсхему (7.20), получим

$$\left|\frac{\eta_{k+1}}{\eta_{k+1/2}}\right| = \left|\frac{1-\sigma_1\, b_1}{1+\sigma_2 b_2}\right|.$$

Тогда отношение амплитуд $\frac{\eta_{k+1}}{\eta_k}$ гармоник будет равно

$$\left|\frac{\eta_{k+1}}{\eta_k}\right| = \left|\frac{\eta_{k+1}}{\eta_{k+1/2}}\right|\left|\frac{\eta_{k+1/2}}{\eta_k}\right| = \left|\frac{1-\sigma_1\, b_1}{1+\sigma_1 b_1}\cdot\frac{1-\sigma_2 b_2}{1+\sigma_2 b_2}\right| < 1. \tag{7.28}$$

Отсюда видно, что схема МПН абсолютна устойчива.

Проанализируем выражение (7.28). Из него видно, что ослабление устойчивости по переменной y в подсхеме (7.19) (уменьшение числителя второго сомножителя) компенсируется большим усилением устойчивости по направлению y в подсхеме (7.20)

(знаменатель второго сомножителя). Аналогично для переменной x. Ослабление устойчивости по переменной x в подсхеме (7.20) (уменьшение числителя первого сомножителя) компенсируется большим усилением устойчивости по направлению x в подсхеме (7.19) (знаменатель первого сомножителя).

Таким образом, в схеме метода переменных направлений ослабление устойчивости по какому-либо направлению, вследствие явной аппроксимации по этому направлению, компенсируется увеличением устойчивости по другому направлению. Из-за этого явления МПН не применяется к трехмерным задачам, поскольку в этом случае он становится условно устойчивым.

К *достоинствам* метода переменных направлений можно отнести высокую точность, поскольку метод имеет второй порядок точности по времени. При малых числах Куранта $\sigma_1 \ll 1$, $\sigma_2 \ll 1$ с помощью МПН в двумерном случае можно тестировать другие численные схемы.

К *недостаткам* можно отнести условную устойчивость при числе пространственных переменных больше двух. Кроме этого, МПН условно устойчив в задачах со смешанными производными уже в двумерном случае. Таким образом, запас устойчивости у метода довольно низок. Можно показать, что в задачах с краевыми условиями 2-го или 3-го рода схема МПН также условно устойчива.

§ 7.3. Метод дробных шагов Н. Н. Яненко

В отличие от МПН, метод дробных шагов (МДШ) использует только неявные конечно-разностные операторы, что делает его абсолютно устойчивым в задачах, не содержащих смешанных производных. Он обладает довольно значительным запасом устойчивости (сохранение устойчивости при числах Куранта, значительно превышающих единицу) и в задачах со смешанными производными.

Для задачи (7.1)–(7.6) схема МДШ имеет вид

$$\frac{u_{ij}^{k+1/2} - u_{ij}^{k}}{\tau} = \frac{a}{h_1^2}\left(u_{i+1j}^{k+1/2} - 2u_{ij}^{k+1/2} + u_{i-1j}^{k+1/2}\right), \tag{7.29}$$

$$\frac{u_{ij}^{k+1} - u_{ij}^{k+1/2}}{\tau} = \frac{a}{h_2^2}\left(u_{i\,j+1}^{k+1} - 2u_{ij}^{k+1} + u_{i\,j-1}^{k+1}\right). \tag{7.30}$$

С помощью чисто неявной подсхемы (7.29) осуществляются скалярные прогонки в направлении оси x в количестве, равном $J-1$, в результате чего получаем сеточную функцию $u_{ij}^{k+1/2}$. На втором дробном шаге по времени с помощью подсхемы (7.30) осуществляются скалярные прогонки в направлении оси y в количестве, равном $I-1$, в результате чего получаем сеточную функцию u_{ij}^{k+1}. Шаблон схемы МДШ приведен на рис. 7.2. Для определения порядка аппроксимации схемы МДШ запишем ее в следующей операторной форме:

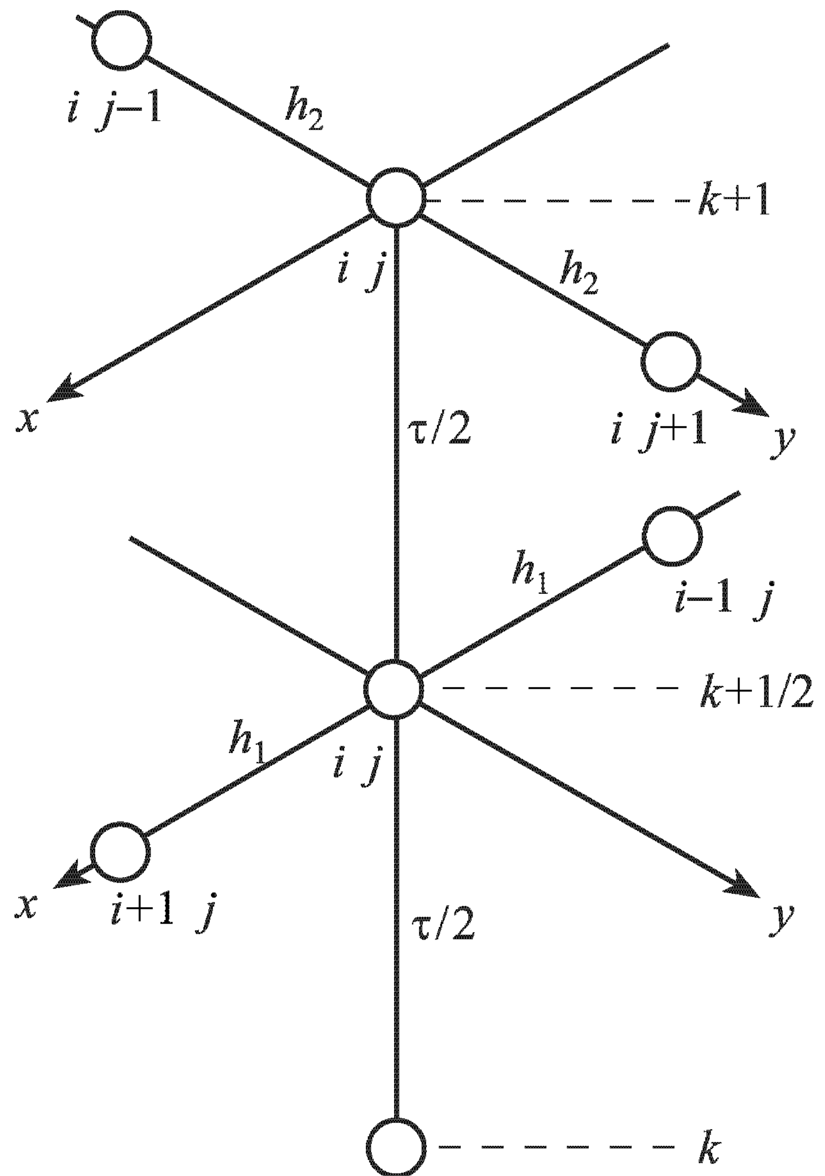


Рис. 7.2. Шаблон схемы метода дробных шагов

$$\frac{u^{k+1/2}-u^k}{\tau}=\Lambda_1 u^{k+1/2},$$

$$\left(E-\tau\Lambda_1\right)u^{k+1/2}=Eu^k,$$

$$\frac{u^{k+1}-u^{k+1/2}}{\tau}=\Lambda_2 u^{k+1},$$

$$\left(E-\tau\Lambda_2\right)u^{k+1}=Eu^{k+1/2}.$$

Исключая здесь сеточную функцию на промежуточном временном слое $t^{k+1/2}=t^k+\dfrac{\tau}{2}$, получим двухслойную схему

$$\left(E-\tau\Lambda_1\right)\left(E-\tau\Lambda_2\right)u^{k+1}=Eu^k,$$

откуда получаем порядок аппроксимации по времени:

$$\left(E-\tau\left(\Lambda_1+\Lambda_2\right)+\tau^2\Lambda_1\Lambda_2\right)\,u^{k+1}=Eu^k,$$

$$\frac{u^{k+1}-u^k}{\tau}=\Lambda_1 u^{k+1}+\Lambda_2 u^{k+1}-\tau\Lambda_1\Lambda_2 u^{k+1}. \tag{7.31}$$

Из (7.31) видно, что схема МДШ (7.29), (7.30) имеет порядок $O(\tau+|h|^2)$, т. е. первый порядок по времени и второй — по переменным x и y.

Для исследования устойчивости схемы МДШ (7.29), (7.30) подставим в нее гармонику (7.27), получим

$$\eta_{k+1/2} - \eta_k = \sigma_1 \eta_{k+1/2}(-b_1), \quad b_1 = 4\sin^2\frac{\lambda_n h_1}{2} > 0;$$

$$\eta_{k+1} - \eta_{k+1/2} = \sigma_2 \eta_{k+1}(-b_2), \quad b_2 = 4\sin^2\frac{\lambda_m h_2}{2} > 0.$$

Тогда отношение амплитуд гармоник равно

$$\left|\frac{\eta_{k+1}}{\eta_k}\right| = \left|\frac{\eta_{k+1}}{\eta_{k+1/2}}\right| \left|\frac{\eta_{k+1/2}}{\eta_k}\right| = \frac{1}{1+\sigma_1 b_1}\cdot\frac{1}{1+\sigma_2 b_2} < 1,$$

т. е. схема метода дробных шагов абсолютно устойчива.

Достоинства схемы МДШ: 1) проста в алгоритмизации и программировании; 2) абсолютно устойчива с большим запасом устойчивости даже для задач, содержащих смешанные производные.

Недостатки: 1) на каждом дробном шаге достигается частичная аппроксимация, полная аппроксимация достигается на последнем дробном шаге; 2) имеет первый порядок точности по времени; 3) в задачах со смешанными производными для устойчивости МДШ на коэффициенты накладываются жесткие ограничения, при невыполнении которых схема становится условно устойчивой.

§ 7.4. Метод переменных направлений с экстраполяцией В. Ф. Формалева

Как видно из изложенных выше конечно-разностных схем для многомерных задач математической физики, конечно-разностная схема может быть либо абсолютно устойчивой, но не экономичной (метод матричной прогонки), либо экономичной, но условно устойчивой, как схема МПН. Желание получить схемы экономичные и одновременно абсолютно устойчивые с большим запасом устойчивости приводит к необходимости использовать апостериорную информацию о сеточной функции, которая получается в процессе счета. К таким схемам относится схема метода переменных направлений с экстраполяцией (МПНЭ). Эта схема использует распределение сеточной функции в левом (от расчетного) пространственном сечении, уже полученное на верхнем временном слое, т. е. неявно. В правом пространственном сечении

значения сеточной функции в первом приближении и с контролируемой точностью определяются линейной экстраполяцией по времени с нижних временных слоев на верхний. Затем эти значения уточняются с помощью скалярных прогонок.

Рассмотрим задачу, аналогичную задаче (7.1)–(7.6), в которой дифференциальное уравнение (7.1) содержит смешанные производные (наличие последних не обязательно):

$$\frac{\partial u}{\partial t} = a_{11}\frac{\partial^2 u}{\partial x^2} + 2a_{12}\frac{\partial^2 u}{\partial x \partial y} + a_{22}\frac{\partial^2 u}{\partial y^2},$$

$$x \in (0;\, \ell_1), \quad y \in (0;\, \ell_2), \quad t > 0. \tag{7.32}$$

На сетке (7.7) аппроксимируем это дифференциальное уравнение с помощью следующей схемы:

$$\frac{u_{ij}^{k+1/2} - u_{ij}^{k}}{\tau/2} = \frac{a_{11}}{h_1^2}\left(u_{i+1j}^{k+1/2} - 2u_{ij}^{k+1/2} + u_{i-1j}^{k+1/2}\right) +$$

$$+ \frac{2a_{12}}{4h_1h_2}\left(\tilde{u}_{i+1j+1}^{k+1/2} - u_{i+1j-1}^{k+1/2} - \tilde{u}_{i-1j+1}^{k+1/2} + u_{i-1j-1}^{k+1/2}\right) +$$

$$+ \frac{a_{22}}{h_2^2}\left(\tilde{u}_{ij+1}^{k+1/2} - 2u_{ij}^{k+1/2} + u_{ij-1}^{k+1/2}\right), \tag{7.33}$$

$$\frac{u_{ij}^{k+1} - u_{ij}^{k+1/2}}{\tau/2} = \frac{a_{11}}{h_1^2}\left(\tilde{u}_{i+1j}^{k+1} - 2u_{ij}^{k+1} + u_{i-1j}^{k+1}\right) +$$

$$+ \frac{2a_{12}}{4h_1h_2}\left(\tilde{u}_{i+1j+1}^{k+1} - \tilde{u}_{i+1\,j-1}^{k+1} - u_{i-1j+1}^{k+1} + u_{i-1j-1}^{k+1}\right) +$$

$$+ \frac{a_{22}}{h_2^2}\left(u_{ij+1}^{k+1} - 2u_{ij}^{k+1} + u_{ij-1}^{k+1}\right). \tag{7.34}$$

Здесь значения сеточной функции $\tilde{u}_{ij}$, помеченные волнистой чертой, определяются с помощью линейной экстраполяции:

$$\tilde{u}_{m\,j+1}^{k+1/2} = 2u_{mj+1}^{k} - u_{mj+1}^{k-1/2}, \quad m = i-1,\, i,\, i+1,$$

$$\tilde{u}_{i+1\,m}^{k+1} = 2u_{i+1m}^{k+1/2} - u_{i+1\,m}^{k}, \quad m = j-1,\, j,\, j+1.$$

Шаблон схемы МПНЭ представлен на рис. 7.3

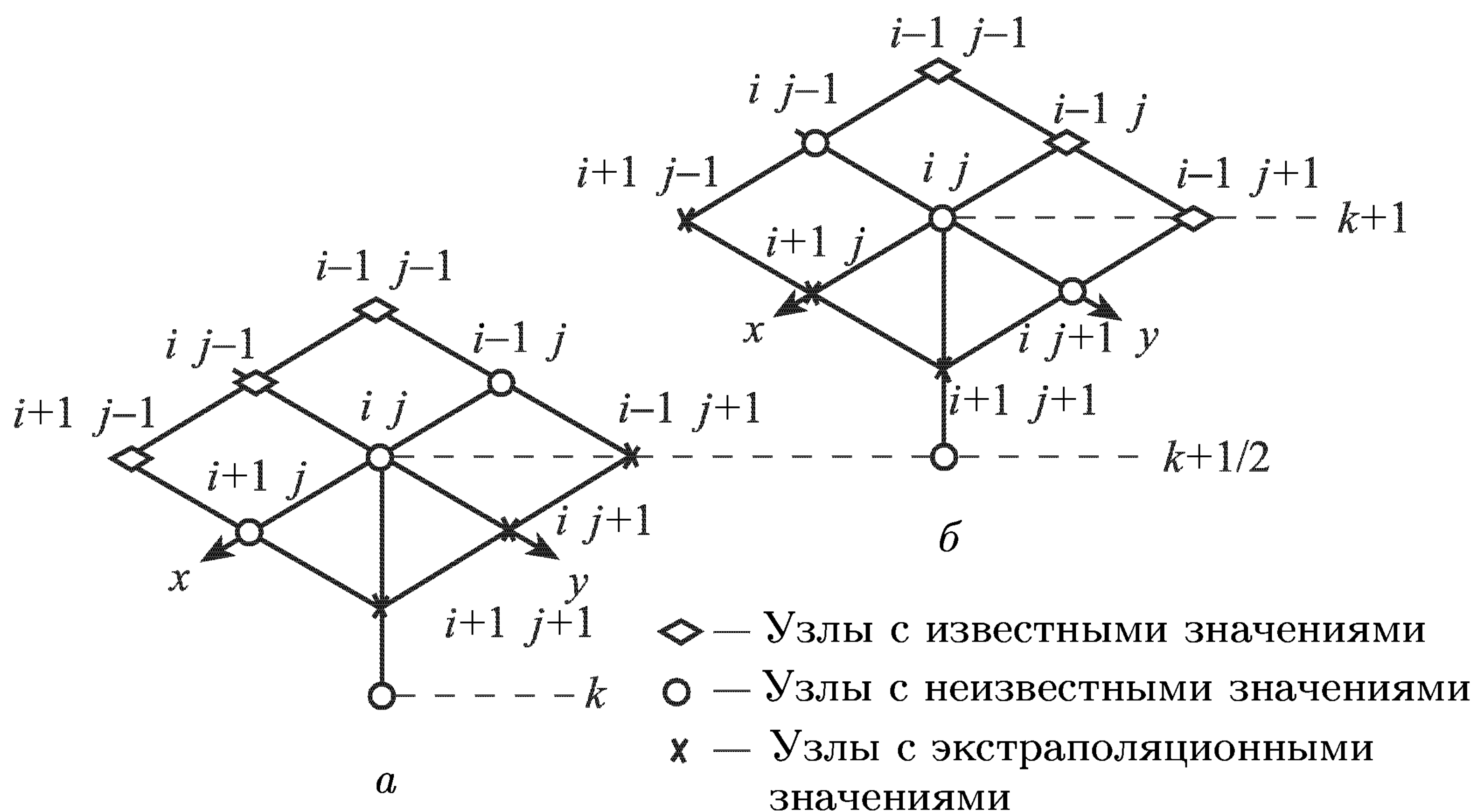


Рис. 7.3. Шаблон схемы метода переменных направлений с экстраполяцией: *а* — подсхема (7.33); *б* — подсхема (7.34)

В подсхеме (7.33) значения $u_{i-1}^{k+1/2}$, $u_{ij}^{k+1/2}$, $u_{i+1\,j}^{k+1/2}$ являются искомыми, определяемыми из скалярных прогонок в направлении переменной x , значения $u_{i-1\,j-1}^{k+1/2}$, $u_{ij-1}^{k+1/2}$, $u_{i+1\,j-1}^{k+1/2}$ уже известны на верхнем полуслое из прогонки вдоль координатной линии $y_{j-1} = (j-1)h_2$, а значения $\widetilde{u}_{i-1j+1}^{k+1/2}$, $\widetilde{u}_{ij+1}^{k+1/2}$, $\widetilde{u}_{i+1j+1}^{k+1/2}$ с порядком $O(\tau^2)$ определяются экстраполяцией по распределениям функции на двух предыдущих временных полуслоях. При этом все конечно-разностные операторы по пространственным переменным, за исключением оператора по переменной x, переводятся в правые части (хотя они и являются практически полностью неявными).

В подсхеме (7.34) значения u_{ij-1}^{k+1}, u_{ij}^{k+1} , u_{ij+1}^{k+1} являются искомыми, определяемыми из скалярных прогонок вдоль переменной y, значения u_{i-1j-1}^{k+1}, u_{i-1j}^{k+1}, u_{i-1j+1}^{k+1} известны как значения сеточной функции в левом пространственном сечении, а значения $\widetilde{u}_{i+1j-1}^{k+1}$, $\widetilde{u}_{i+1j}^{k+1}$, $\widetilde{u}_{i+1j+1}^{k+1}$ с порядком $O(\tau^2)$ определяются экстраполяцией по двум предыдущим временным полуслоям. При этом все пространственные операторы, за исключением оператора по переменной y, переводятся в правые части, хотя они практически являются полностью неявными.

7.4.1. Аппроксимация. Для анализа аппроксимационных свойств схемы (7.33), (7.34) прибавим и вычтем в подсхеме (7.33) выражения $\frac{a_{22}}{h_2^2}u_{ij+1}^{k+1/2}$, $\frac{2a_{12}}{4h_1h_2}\left(u_{i+1j+1}^{k+1/2} - u_{i-1j+1}^{k+1/2}\right)$, а в подсхеме (7.34) — выражения $\frac{a_{11}}{h_1^2}u_{i+1j}^{k+1}$, $\frac{2a_{12}}{4h_1h_2}\left(u_{i+1\,j+1}^{k+1} - u_{i+1\,j-1}^{k+1}\right)$, получим эквивалентную схему

$$\frac{u^{k+1/2} - u^k}{\tau/2} = \Lambda u^{k+1/2} + \left(\Gamma_{22} + 2\Gamma_{12}^I\right)\widetilde{u}^{k+1/2}, \tag{7.35}$$

$$\frac{u^{k+1} - u^{k+1/2}}{\tau/2} = \Lambda u^{k+1} + \left(\Gamma_{11} + 2\Gamma_{12}^{II}\right)\widetilde{u}^{k+1}, \tag{7.36}$$

где $\Lambda = \Lambda_{11} + 2\Lambda_{12} + \Lambda_{22}$,

$$\Lambda_{11}u = \frac{a_{11}}{h_1^2}\left(u_{i+1j} - 2u_{ij} + u_{i-1j}\right),$$

$$\Lambda_{12}u = \frac{2a_{12}}{4h_1h_2}\left(u_{i+1j+1} - u_{i+1j-1} - u_{i-1j+1} + u_{i-1j-1}\right),$$

$$\Lambda_{22}u = \frac{a_{22}}{h_2^2}\left(u_{ij+1} - 2u_{ij} + u_{ij-1}\right),$$

$$\Gamma_{22}\widetilde{u}^{k+1/2} = -\frac{a_{22}}{h_2^2}\left(u_{ij+1}^{k+1/2} - \widetilde{u}_{ij+1}^{k+1/2}\right),$$

$$\Gamma_{12}^{\mathrm{I}}\widetilde{u}^{k+1/2} =$$

$$= -\frac{2a_{12}}{4h_1h_2}\left[\left(u_{i+1j+1}^{k+1/2} - \widetilde{u}_{i+1j+1}^{k+1/2}\right) - \left(u_{i-1j+1}^{k+1/2} - \widetilde{u}_{i-1j+1}^{k+1/2}\right)\right],$$

$$\Gamma_{11}\widetilde{u}^{k+1} = -\frac{a_{11}}{h_1^2}\left(u_{i+1j}^{k+1} - \widetilde{u}_{i+1j}^{k+1}\right),$$

$$\Gamma_{12}^{\mathrm{II}}\widetilde{u}^{k+1} = -\frac{2a_{12}}{4h_1h_2}[(u_{i+1j+1}^{k+1} - \widetilde{u}_{i+1j+1}^{k+1}) - (u_{i+1j-1}^{k+1} - \widetilde{u}_{i+1j-1}^{k+1})].$$

Справедлива следующая теорема.

Теорема 7.1. *Пусть решение задачи* (7.32), (7.2)–(7.6) $u(x,y,t) \in C_4^2\left(\overline{G}_T\right)$, где $\overline{G}_T == \overline{G} \times [0,T]$, $t \in [0,T]$, C_n^m — *класс функций, m раз непрерывно дифференцируемых по t и n раз — по x, y. Тогда схема* (7.33), (7.34) *и, следовательно, эквивалентная ей схема* (7.35), (7.36) *аппроксимирует на точном*

решении дифференциальную задачу (7.32), (7.2)–(7.6) *с порядком* $O\left(\tau + |h|^2 + \tau\,(h_1 + h_2)\right)$, *где* $|h|^2 = h_1^2 + h_2^2$.

Доказательство. Действительно, рассматривая «осколочные» операторы Γ_{22}, Γ_{12}^{I}, Γ_{11}, $\Gamma_{12}^{\mathrm{II}}$, можно заметить, что выполняются следующие тождества :

$$u_{ij+1}^{k+1/2} - u_{ij+1}^{k} = \\ = \left(u_{ij}^{k+1/2} - u_{ij}^{k}\right) + \frac{u_{ij+1}^{k+1/2} - u_{ij}^{k+1/2}}{h_2} h_2 - \frac{u_{ij+1}^{k} - u_{ij}^{k}}{h_2} h_2,$$

$$u_{ij+1}^{k} - u_{ij+1}^{k-1/2} = \\ = \left(u_{ij}^{k} - u_{ij}^{k-1/2}\right) + \frac{u_{ij+1}^{k} - u_{ij}^{k}}{h_2} h_2 - \frac{u_{ij+1}^{k-1/2} - u_{ij}^{k-1/2}}{h_2} h_2,$$

$$u_{i+1j}^{k+1} - u_{i+1j}^{k+1/2} = \\ = \left(u_{ij}^{k+1} - u_{ij}^{k+1/2}\right) + \frac{u_{i+1j}^{k+1} - u_{ij}^{k+1}}{h_1} h_1 - \frac{u_{i+1j}^{k+1/2} - u_{ij}^{k+1/2}}{h_1} h_1,$$

$$u_{i+1j}^{k+1/2} - u_{i+1j}^{k} = \\ = \left(u_{ij}^{k+1/2} - u_{ij}^{k}\right) + \frac{u_{i+1j}^{k+1/2} - u_{ij}^{k+1/2}}{h_1} h_1 - \frac{u_{i+1j}^{k} + u_{ij}^{k}}{h_1} h_1,$$

$$\frac{u^{k+1/2} - u^{k}}{\tau/2} = \Lambda u^{k+1/2}, \qquad \frac{u^{k+1} - u^{k+1/2}}{\tau/2} = \Lambda u^{k+1},$$

$$u_y = \frac{(u_{ij+1} - u_{ij})}{h_2}, \qquad u_x = \frac{(u_{i+1j} - u_{ij})}{h_1}, \qquad u_{\overline{y}} = \frac{(u_{ij} - u_{ij-1})}{h_2},$$

$$u_{\overline{x}} = \frac{(u_{ij} - u_{i-1j})}{h_1}, \qquad u_{tt}^{k} = \frac{\left(u_{ij}^{k+1/2} - 2u_{ij}^{k} + u_{ij}^{k-1/2}\right)}{\tau^2/4},$$

$$\sigma_{i\,i} = \frac{a_{i\,i}\tau}{2h_i^2}, \quad i = 1, 2, \quad \sigma_{12} = \frac{a_{1\,2}\tau}{8h_1h_2}.$$

Тогда справедлива следующая цепочка равенств:

$$\Gamma_{2\,2}\widetilde{u}^{k+1/2} = -\frac{a_{2\,2}}{h_2^2}\left(u_{ij+1}^{k+1/2} - \widetilde{u}_{ij+1}^{k+1/2}\right) =$$

$$= -\frac{a_{22}}{h_2^2}\left(u_{ij+1}^{k+1/2} - 2u_{ij+1}^{k} + u_{ij+1}^{k-1/2}\right) =$$

$$= -\frac{a_{22}}{h_2^2}\left[\left(u_{ij+1}^{k+1/2} - u_{ij+1}^{k}\right) - \left(u_{ij+1}^{k} - u_{ij+1}^{k-1/2}\right)\right] =$$

$$= -\frac{a_{22}}{h_2^2}\left[\left(u_{ij}^{k+1/2} - u_{ij}^{k}\right) - \left(u_{ij}^{k} - u_{ij}^{k-1/2}\right) + \right.$$

$$\left. + \left(u_y^{k+1/2} - 2u_y^{k} + u_y^{k-1/2}\right)_{i\,j} h_2\right] =$$

$$= -\frac{a_{22}}{h_2^2}\left(\frac{\tau}{2}\Lambda u^{k+1/2} - \frac{\tau}{2}\Lambda u^{k} + u_{tty}^{k}\frac{\tau^2 h_2}{4}\right) =$$

$$= -\frac{a_{22}}{h_2^2}\left[\frac{\tau}{2}\Lambda\left(u^{k+1/2} - u^{k}\right) + u_{tty}^{k}\frac{\tau^2 h_2}{4}\right] =$$

$$= -\frac{a_{22}}{h_2^2}\left[\frac{\tau^2}{4}\Lambda\left(\frac{u^{k+1/2} - u^{k}}{\tau/2}\right) + u_{tty}^{k}\frac{\tau^2 h_2}{4}\right] =$$

$$= -\frac{a_{22}\tau^2}{4h_2^2}\Lambda^2 u^{k+1/2} - \sigma_{22}u_{tty}^{k}\cdot\frac{\tau}{4}h_2 =$$

$$= -\frac{a_{22}\tau^2}{4h_2^2}\Lambda^2 u^{k+1/2} + O\left(\tau h_2\right).$$

Таким образом,

$$\Gamma_{22}\widetilde{u}^{k+1/2} = -\frac{\sigma_{22}\tau}{2}\Lambda^2 u^{k+1/2} + O\left(\tau h_2\right). \tag{7.37}$$

Аналогично,

$$\Gamma_{11}\widetilde{u}^{k+1} = -\frac{\sigma_{11}\tau}{2}\Lambda^2 u^{k+1} + O\left(\tau h_1\right); \tag{7.38}$$

$$\Gamma_{12}^{\mathrm{I}}\widetilde{u}^{k+1/2} =$$

$$= -\frac{2a_{12}}{4h_1h_2}\left[\left(u_{i+1j+1}^{k+1/2} - \widetilde{u}_{i+1j+1}^{k+1/2}\right) - \left(u_{i-1j+1}^{k+1/2} - \widetilde{u}_{i-1j+1}^{k+1/2}\right)\right] =$$

$$= -\frac{2a_{12}}{4h_1h_2}\left[\left(u_{i+1j+1}^{k+1/2} - 2u_{i+1j+1}^{k} + u_{i+1j+1}^{k-1/2}\right) - \right.$$

$$\left. - \left(u_{i-1j+1}^{k+1/2} - 2u_{i-1j+1}^{k} + u_{i-1j+1}^{k-1/2}\right)\right] =$$

$$= -\frac{2a_{12}}{4h_1h_2}\left[\left(u^k_{ttx} + u^k_{tt\overline{x}}\right)\frac{\tau^2 h_1}{4} + \left(u^k_{ttxy} - u^k_{tt\overline{x}y}\right)\frac{\tau^2 h_2}{4}h_1\right] =$$

$$= -2\sigma_{12}\left(u^k_{ttx} + u^k_{tt\overline{x}}\right)\tau h_1 - \frac{a_{12}}{8}\left(u^k_{ttxy} - u^k_{tt\overline{x}y}\right)\tau^2 =$$

$$= O\left(\tau h_1 + \tau^2\right); \quad (7.39)$$

$$\Gamma^{II}_{12}\widetilde{u}^{k+1} = O\left(\tau h_2 + \tau^2\right). \quad (7.40)$$

Подставляя выражения (7.37)–(7.40) для «осколочных» операторов в схему (7.35), (7.36) , получим

$$\frac{u^{k+1/2} - u^k}{\tau/2} = \Lambda u^{k+1/2} - \frac{\sigma_{22}\tau}{2}\Lambda^2 u^{k+1/2} + $$
$$+ O\left(\tau + |h|^2 + \tau\left(h_1 + h_2\right) + \tau^2\right), \quad (7.41)$$

$$\frac{u^{k+1} - u^{k+1/2}}{\tau/2} = \Lambda u^{k+1} - \frac{\sigma_{11}\tau}{2}\Lambda^2 u^{k+1} + $$
$$+ O\left(\tau + |h|^2 + \tau\left(h_1 + h_2\right) + \tau^2\right). \quad (7.42)$$

Исключение сеточной функции $u^{k+1/2}$ на промежуточном временном полуслое приводит к следующей эквивалентной двухслойной схеме:

$$\frac{u^{k+1} - u^k}{\tau} = \Lambda\left[E - (1+a)\frac{\tau}{4}\Lambda + a\frac{\tau^2}{8}\Lambda^2 - b\frac{\tau^3}{16}\Lambda^3\right]u^{k+1} + $$
$$+ O\left(\tau + |h|^2 + \tau\left(h_1 + h_2\right) + \tau^2\right), \quad (7.43)$$

где $a = \sigma_{11} + \sigma_{22} > 0$, $b = \sigma_{11}\sigma_{22} > 0$.

Из (7.43) следует аппроксимация с порядком $O\Big(\tau + |h|^2 + \tau\left(h_1 + h_2\right)\Big)$. Теорема доказана.

7.4.2. Устойчивость. Будем исследовать устойчивость схемы (7.33), (7.34) методом энергетических неравенств.

Имеет место следующая теорема.

Теорема 7.2. *Пусть выполнены условия* $a_{11} > 0$, $a_{22} > 0$. *Тогда схема* (7.33), (7.34) *абсолютна устойчива по начальным условиям.*

Для доказательства этой теоремы докажем следующую лемму.

Лемма 7.1. *Оператор*

$$C = \left[-(1+\sigma_{11}+\sigma_{22})\frac{\tau}{4}\Lambda + (\sigma_{11}+\sigma_{22})\frac{\tau^2}{8}\Lambda^2 - \sigma_{11}\sigma_{22}\frac{\tau^3}{16}\Lambda^3\right]$$

в эквивалентной схеме (7.43) *является положительно определенным и самосопряженным.*

Действительно, этот оператор можно переписать в виде

$$\mathrm{C} = (1+a)\frac{\tau}{4}A + a\frac{\tau^2}{8}A^2 + b\frac{\tau^3}{16}A^3,$$

в котором каждое слагаемое является положительно определенным и самосопряженным оператором, поскольку $-\Lambda = A$ — положительный, самосопряженный оператор, действующий в гильбертовом пространстве H (см. гл. 6). Следовательно, оператор C является положительно определенным, самосопряженным оператором, что доказывает лемму.

Таким образом, (7.43) можно переписать в виде

$$\frac{u^{k+1}-u^k}{\tau} = -A\,(E+C)\,u^{k+1}.$$

Умножая это равенство скалярно на $u_t = \dfrac{\left(u^{k+1}-u^k\right)}{\tau}$ и используя известные тождества

$$u^{k+1} \equiv \frac{\left(u^{k+1}+u^k\right)}{2} + \frac{\left(u^{k+1}-u^k\right)}{2} \equiv \frac{\left(u^{k+1}+u^k\right)}{2} + \frac{\tau u_t}{2},$$

получим следующее энергетическое тождество:

$$\left(\left[E + \frac{\tau}{2}(A+AC)\right]u_t,\, u_t\right) + \frac{1}{2\tau}\left[\left((A+AC)\,u^{k+1},\, u^{k+1}\right) - \right.$$
$$\left. - \left((A+AC)u^k,\, u^k\right)\right] = 0, \quad (7.44)$$

при выводе которого использована положительность и самосопряженность оператора $D = A(E+C)$.

Вводя энергетическое пространство H_D элементов $u,\ v,\ \ldots$ $\ldots \in \mathrm{H}_D$ со скалярным произведением $(u,v)_D = (Du,v)$ и нормой $\|\,u\,\|_D^2 = (Du,u)$, в силу положительности оператора $E + \frac{\tau}{2}(A+AC)$ из (7.44) получаем энергетическое неравенство

$$\left(Du^{k+1},\, u^{k+1}\right) \leqslant \left(Du^k,\, u^k\right),$$

откуда следует принцип максимума

$$\left\|u^{k+1}\right\|_D \leqslant \left\|u^k\right\|_D \leqslant \left\|u^{k-1}\right\|_D \leqslant \ldots \leqslant \left\|u^0\right\|_D = \|\psi(x,y)\|_D,$$

являющийся достаточным признаком устойчивости конечно-разностной схемы (7.33), (7.34), что доказывает теорему.

Поскольку не накладывалось никаких ограничений на сеточные характеристики τ, h_1, h_2, схема (7.33), (7.34) является *абсолютно устойчивой.*

По запасу устойчивости метод МПНЭ превосходит все существующие экономичные методы расщепления для задач как содержащих, так и не содержащих смешанные дифференциальные операторы.

К *достоинствам МПНЭ* можно отнести следующие: 1) экономичность; 2) абсолютную устойчивость; 3) полную (не частичную, как в МДШ) аппроксимацию дифференциального уравнения; 4) применимость к задачам с любой размерностью по пространственным переменным и к задачам, содержащим смешанные дифференциальные операторы; 5) отсутствие ограничений на величину коэффициентов a_{11}, a_{12}, a_{22}, кроме ограничений $a_{11} > 0$, $a_{22} > 0$, $a_{11}a_{22} > a_{12}^2$.

§ 7.5. Схема метода полного расщепления Формалева–Тюкина

Аналогичной схемой, использующей апостериорную информацию, является схема метода Формалева–Тюкина. Эта полностью неявная, но экономичная схема основана на более глубоком расщеплении смешанных дифференциальных операторов по сравнению со схемами, разобранными выше.

Для задачи (7.32), (7.2)–(7.6) на сетке (7.7) эта схема имеет следующий вид [20,21]:

$$\begin{aligned}
\frac{u_{ij}^{k+1/2} - u_{ij}^k}{\tau} &= \frac{a_{11}}{h_1^2}\left(u_{i+1j}^{k+1/2} - 2u_{ij}^{k+1/2} + u_{i-1j}^{k+1/2}\right) + \\
&+ \frac{1-\sigma}{2}\frac{a_{12}}{h_1h_2}\left(u_{ij}^{k+1/2} - u_{ij-1}^{k+1/2} - u_{i-1j}^{k+1/2} + u_{i-1j-1}^{k+1/2}\right) + \\
&+ \frac{1+\sigma}{2}\frac{a_{12}}{h_1h_2}\left(u_{i+1j}^{k+1/2} - u_{i+1j-1}^{k+1/2} - u_{ij}^{k+1/2} + u_{ij-1}^{k+1/2}\right); \qquad (7.45)
\end{aligned}$$

$$\frac{u_{ij}^{k+1} - u_{ij}^{k+1/2}}{\tau} = \frac{a_{22}}{h_2^2}\left(u_{ij+1}^{k+1} - 2u_{ij}^{k+1} + u_{ij-1}^{k+1}\right) + \\ + \frac{1-\sigma}{2}\frac{a_{12}}{h_1 h_2}\left(u_{i+1j+1}^{k+1} - u_{i+1j}^{k+1} - u_{ij+1}^{k+1} + u_{ij}^{k+1}\right) + \\ + \frac{1+\sigma}{2}\frac{a_{12}}{h_1 h_2}\left(u_{ij+1}^{k+1} - u_{ij}^{k+1} - u_{i-1j+1}^{k+1} + u_{i-1j}^{k+1}\right), \quad (7.46)$$

причем $\sigma = -1$, если $a_{12} < 0$, $\sigma = 1$, если $a_{12} > 0$. Значения $a_{11} > 0$, $a_{22} > 0$, $a_{11} \cdot a_{22} - a_{12}^2 > 0$. Коэффициент σ введен для того, чтобы анализ устойчивости не зависел от знака коэффициента a_{12}.

Шаблоны схемы (7.45), (7.46) приведены на рис. 7.4.

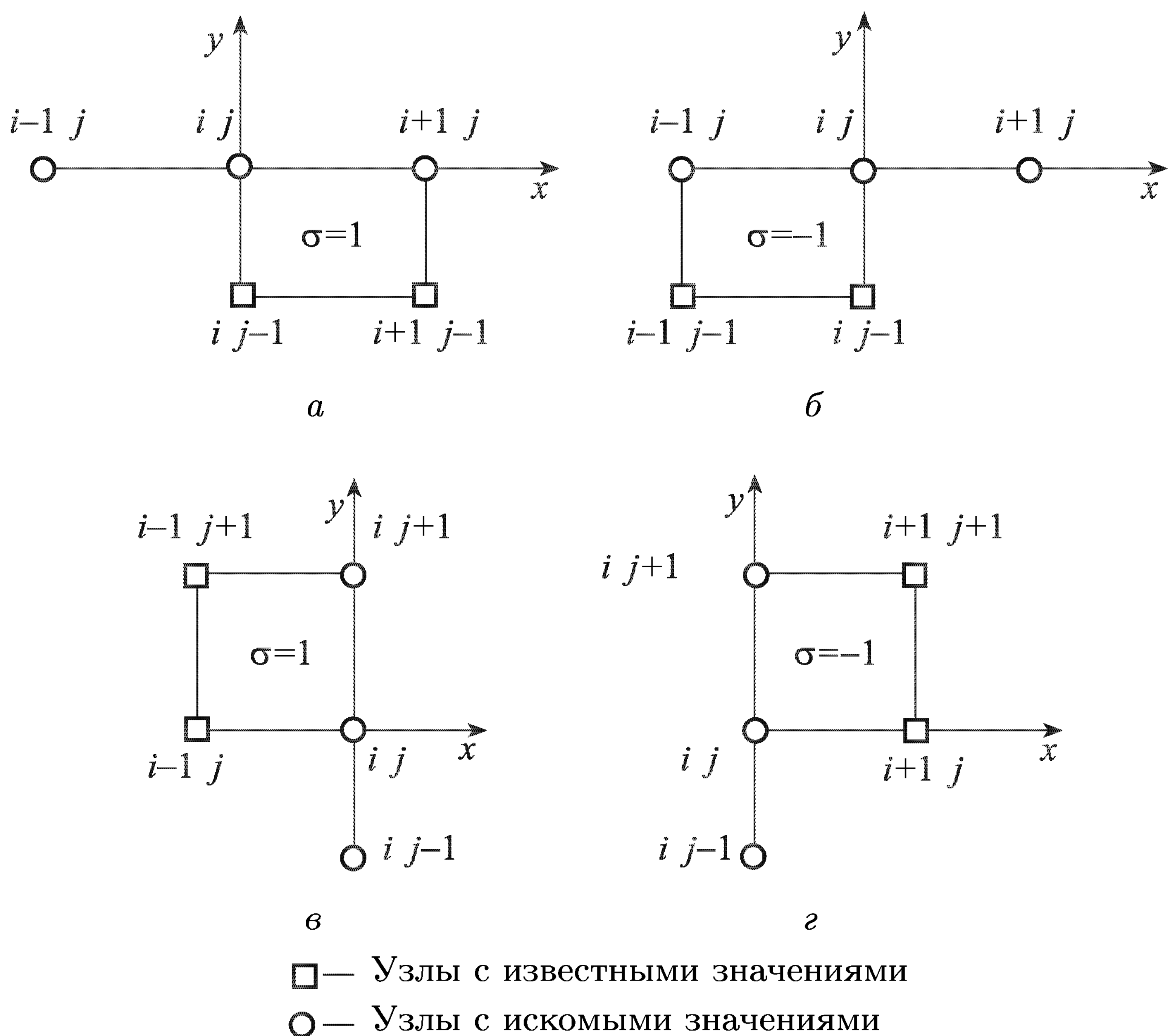


Рис. 7.4. Шаблоны схемы полного расщепления: *а*, *б* — подсхема (7.45); *в*, *г* — подсхема (7.46)

Подсхема (7.45) реализуется с помощью скалярных прогонок в направлении координатной оси Ox, при этом значения сеточной функции в нижнем $(j-1)$-м сечении уже известны на

верхнем временном полуслое. В подсхеме индекс i изменяется от 1 до $I-1$, а индекс j — от 1 до $J-1$.

Подсхема (7.46) реализуется с помощью скалярных прогонок в направлении координатной оси Oy, при этом значения сеточной функции в левом $(i-1)$-м сечении (шаблон 7.4 *в*) или в правом $(i+1)$-м сечении (шаблон 7.4 *г*) уже известны на верхнем временном слое.

В подсхеме индекс j изменятся от 1 до $J-1$, а индекс i — от 1 до $I-1$ (шаблон 7.4 *в*) или от $I-1$ до 1 (шаблон 7.4 *г*).

Таким образом, схема полного расщепления (7.45), (7.46) является экономичной, поскольку реализуется с помощью скалярных прогонок, несмотря на то что дифференциальное уравнение содержит смешанные производные, и полностью неявной, вследствие чего можно ожидать, что схема является абсолютно устойчивой.

Как и схема метода дробных шагов Н. Н. Яненко, схема (7.45), (7.46) обладает частичной аппроксимацией на каждом дробном шаге и полной аппроксимацией — на целом шаге по времени.

Разлагая значения сеточной функции в окрестности узла $\left(x_i,\, y_j,\, t^k\right)$ на точном решении в ряд Тейлора до второй производной по времени и до четвертой производной по пространственным переменным, можно показать, что порядок полной аппроксимации в схеме (7.45), (7.46) составляет $O\left(\tau + h_1^2 + h_2^2\right)$.

Аналогично схеме (7.33), (7.34) с помощью метода энергетических неравенств можно показать абсолютную устойчивость схемы (7.45), (7.46) полного расщепления.

Таким образом, за счет более глубокого расщепления смешанного конечно-разностного оператора в рассматриваемой схеме удалось смешанные производные аппроксимировать неявно, чего не достигалось в классических схемах, рассмотренных выше.

В соответствии с этим достоинствами схемы полного расщепления являются следующие:

— экономичность;

— неявная аппроксимация всех дифференциальных операторов, включая смешанные и, как следствие, абсолютная устойчивость;

— применимость схемы к задачам любой размерности по пространственным переменным;

— отсутствие ограничений на величину коэффициентов a_{11}, $a_{12} = a_{21}$, a_{22}, за исключением естественных ограничений вида $a_{11} > 0$, $a_{22} > 0$, $a_{11}a_{22} > a_{12}^2$.

§ 7.6. Методы расщепления численного решения эллиптических задач

Для стационарных многомерных задач математической физики искомая функция не зависит от времени и, следовательно, уравнение (7.1) в задаче (7.1)–(7.6) становится уравнением эллиптического типа $\left(\text{нет производной } \dfrac{\partial u}{\partial t}\right)$, т. е. уравнением Лапласа или Пуассона, а поскольку и начальное условие (7.6) отсутствует, то рассмотренные выше методы приходится несколько видоизменить.

Однако если при решении задач для уравнений Лапласа или Пуассона используется метод установления (см. § 6.8), то стационарное уравнение Лапласа или Пуассона трансформируется в нестационарное уравнение (7.1), являющееся уже уравнением параболического типа, с введением однородного начального условия (7.6), т. е. $\psi(x, y) \equiv 0$. В этом случае все вышерассмотренные методы применяются без изменения.

Методы расщепления напрямую можно применять также и к решению стационарных задач, *заменив номер временного слоя номером итерации*. При этом в соответствующем конечно-разностном методе конечно-разностная производная по времени $\dfrac{u_{ij}^{k+1} - u_{ij}^{k}}{\tau}$ будет отсутствовать.

В качестве начального приближения на нулевой итерации можно использовать линейную интерполяцию краевых условий (7.2)–(7.5) так, как это делалось в разностно-итерационном методе Либмана (п. 6.2.3).

§ 7.7. Методы решения задач для уравнений гиперболического типа

Одним из наиболее мощных методов решения задач для уравнений гиперболического типа является метод характеристик, применимый как к задачам с начальными условиями, так и к краевым задачам и начально-краевым задачам [22], [23].

В п. 6.1.2 задача для волнового уравнения сводилась к аналогичной задаче для гиперболической системы, которая могла быть решена численно конечно-разностным методом. В более общей постановке квазилинейные гиперболические системы описывают газодинамические течения с вектором скорости, имеющим компоненты $u(x, y)$ и $v(x, y)$ в плоской системе координат xOy.

7.7.1. Метод характеристик решения квазилинейных гиперболических систем. Рассмотрим следующую систему дифференциальных уравнений в частных производных.

$$\begin{cases} a_{11}u_x + a_{12}v_x + b_{11}u_y + b_{12}v_y = c_1, \\ a_{21}u_x + a_{22}v_x + b_{21}u_y + b_{22}v_y = c_2, \end{cases} \tag{7.47}$$

где u, v — искомые функции переменных x, y ; a_{ij}, b_{ij}, i, $j = \overline{1,2}$ — коэффициенты, c_i — правые части, а u_x, v_x, u_y, v_y — частные производные первого порядка. В общем случае a_{ij}, b_{ij}, i, $j = \overline{1,2}$, c_i, $i = \overline{1,2}$ являются функциями x, y, u, v. Такие системы называют *квазилинейными* (линейными относительно первых производных u_x, v_x, u_y, v_y).

Пусть в плоскости xOy задана кривая Γ уравнением $y = y(x)$ и пусть в точках этой кривой заданы значения функций $u(x, y(x))$, $v(x, y(x))$ (рис. 7.5).

Поставим задачу: по значениям решения $u(x, y)$, $v(x, y)$ на кривой Γ найти на той же кривой значения частных производных u_x, u_y, v_x, v_y, используя уравнения (7.47). Если эту задачу удастся решить, то с точностью до малых второго порядка можно определить решения $u(x, y)$, $v(x, y)$ в точках M', близких к точкам кривой Γ, используя разложения

$$\begin{cases} u_{M'} = u_M + u_{xM}\Delta x + u_{yM}\Delta y, \\ v_{M'} = v_M + v_{xM}\Delta x + v_{yM}\Delta y. \end{cases} \tag{7.48}$$

Для этого к системе (7.47) присоединим следующие дифференциальные соотношения на кривой Γ:

$$\left.\frac{du}{dx}\right|_{\Gamma} = u_x + u_y \cdot y', \quad \left.\frac{dv}{dx}\right|_{\Gamma} = v_x + v_y \cdot y'. \tag{7.49}$$

Поскольку на кривой Γ функции $u(x, y(x))\big|_{\Gamma}$ и $v(x, y(x))\big|_{\Gamma}$ являются функциями одной переменной, то полные производные

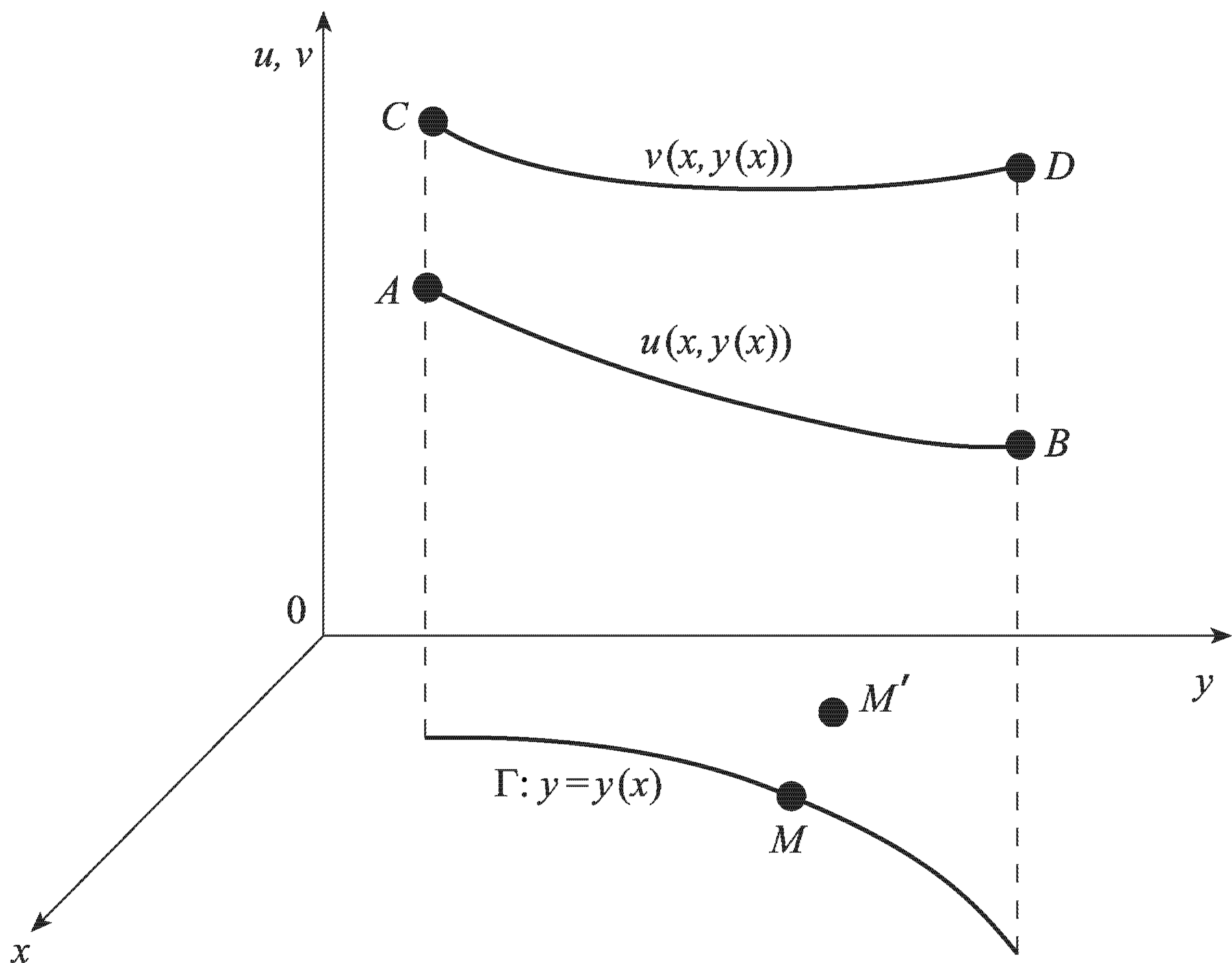


Рис. 7.5. К методу характеристик

(7.49) существуют. При этом $\left.\dfrac{du}{dx}\right|_\Gamma$ и $\left.\dfrac{dv}{dx}\right|_\Gamma$ известны, так как известны $u|_\Gamma$ и $v\,|_\Gamma$. Решаем СЛАУ (7.47), (7.49) относительно u_x, v_x, u_y, v_y, для чего из (7.49) определим u_x, v_x (индекс Γ опущен):

$$u_x = \frac{du}{dx} - u_y y'; \quad v_x = \frac{dv}{dx} - v_y y',$$

подставляя которые в (7.47), получим систему относительно u_y, v_y на кривой Γ:

$$\begin{cases} (b_{11} - a_{11}y')u_y + (b_{12} - a_{12}y')v_y = c_1 - a_{11}\dfrac{du}{dx} - a_{12}\dfrac{dv}{dx}, \\ (b_{21} - a_{21}y')u_y + (b_{22} - a_{22}y')v_y = c_2 - a_{21}\dfrac{du}{dx} - a_{22}\dfrac{dv}{dx}. \end{cases} \tag{7.50}$$

Для определения u_y, v_y составим определитель матрицы системы (7.50):

$$\Delta = \begin{vmatrix} b_{11} - a_{11}y' & b_{12} - a_{12}y' \\ b_{21} - a_{21}y' & b_{22} - a_{22}y' \end{vmatrix}. \tag{7.51}$$

Тогда возможны следующие два случая:

1) если $\Delta \neq 0$ во всех точках кривой Γ, то существует единственное решение u_y, v_y в этих точках, а следовательно и единственное решение u_x, v_x;

2) если $\Delta = 0$, то решение (если оно существует), неединственное.

Будем предполагать, что *решение СЛАУ* (7.50) *существует, и, желая получить не только* u_y, v_y, *но и геометрические места точек* (x, y), *в которых ищется решение, будем рассматривать второй случай*:

$$\Delta = \begin{vmatrix} b_{11} - a_{11}\lambda & b_{12} - a_{12}\lambda \\ b_{21} - a_{21}\lambda & b_{22} - a_{22}\lambda \end{vmatrix} = 0, \qquad (7.52)$$

где $\lambda = y'(x)$.

Раскрывая определитель в (7.52), получаем квадратное уравнение, которое называется *характеристическим*:

$$a\lambda^2 - 2b\lambda + c = 0.$$

Его решением будут следующие два обыкновенных дифференциальных уравнения:

$$\lambda_{1,2} = \frac{b \pm \sqrt{b^2 - ac}}{a} \equiv \left.\frac{dy}{dx}\right|_{1.2}. \qquad (7.53)$$

Общие интегралы этих уравнений на плоскости xOy дадут два семейства интегральных кривых, называемых *характеристиками 1-го и 2-го семейства*, а уравнения (7.53) называются *дифференциальными уравнениями направления характеристик* (*или просто дифференциальными уравнениями характеристик*).

Эти уравнения нелинейные, поскольку коэффициенты зависят не только от x и y, но и от искомых функций $u(x, y)$, $v(x, y)$. Поэтому проинтегрировать их можно численно.

Из курса математической физики известна *классификация системы уравнений* (7.47) в зависимости от знака дискриминанта $b^2 - ac$.

Если $b^2 - ac > 0$, *то система* (7.47) *является системой гиперболического типа, характеристическое уравнение дает два семейства вещественных характеристик с различными направлениями характеристик* $\lambda_1 \neq \lambda_2$.

Если $b^2 - ac = 0$, *то система* (7.47) — *система параболического типа, и при* $b^2 - ac < 0$ *система* (7.47) — *система эллиптического типа.*

Будем рассматривать гиперболическую систему (7.47) *с различными и действительными корнями характеристического уравнения* $\lambda_1 = y'(x)$, $\lambda_2 = y'(x)$.

Таким образом, дифференциальные уравнения направления характеристик (7.53) определят геометрические места точек на плоскости xOy в окрестности кривой Γ, в которых должно быть определено решение $u(x, y)$, $v(x, y)$ исходной задачи (7.47), для чего необходимо составить еще два уравнения относительно $u(x, y)$, $v(x, y)$.

Для их составления рассмотрим определители второго порядка Δ_1, Δ_2 системы (7.50), первый из которых Δ_1 образован из определителя Δ (7.51) заменой первого столбца на столбец правых частей, а второй Δ_2 — из определителя Δ заменой второго столбца на столбец правых частей.

Поскольку сделано *предположение о существовании решения* СЛАУ (7.50), то в соответствии с теоремой Кронекера–Капелли ранг матрицы СЛАУ равен рангу расширенной матрицы (для СЛАУ (7.50) ранг матрицы СЛАУ равен единице, т. е. меньше порядка матрицы, равного двум). Следовательно, поскольку $\Delta = 0$, то в условиях предположения о наличии решения системы (7.50) и $\Delta_1 = \Delta_2 = 0$. Поэтому достаточно рассмотреть одно из уравнений $\Delta_1 = 0$ или $\Delta_2 = 0$, например $\Delta_1 = 0$. Получим

$$\Delta_1 = \begin{vmatrix} c_1 - a_{11}\dfrac{du}{dx} - a_{12}\dfrac{dv}{dx} & b_{12} - a_{12}\lambda_i \\ c_2 - a_{21}\dfrac{du}{dx} - a_{22}\dfrac{dv}{dx} & b_{22} - a_{22}\lambda_i \end{vmatrix} = 0, \quad i = \overline{1,2}. \quad (7.54)$$

Эти два уравнения называются *дифференциальными соотношениями на характеристиках*. После умножения (7.54) на dx и раскрытия определителя можно записать

$$\begin{cases} (A\lambda_1 + B)du + Cdv + Mdx + Ndy = 0, \\ (A\lambda_2 + B)du + Cdv + Mdx + Ndy = 0, \end{cases} \quad (7.55)$$

где

$$A = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad C = \det\begin{pmatrix} b_{12} & a_{12} \\ b_{22} & a_{22} \end{pmatrix},$$

$$B = \det\begin{pmatrix} b_{12} & a_{11} \\ b_{22} & a_{21} \end{pmatrix},$$

$$M = \det\begin{pmatrix} c_1 & b_{12} \\ c_2 & b_{22} \end{pmatrix}, \quad N = \det\begin{pmatrix} a_{12} & c_1 \\ a_{22} & c_2 \end{pmatrix}.$$

Таким образом, решив систему ОДУ (7.55) относительно u, v, ответим на поставленный в задаче вопрос.

Поскольку система (7.47) квазилинейная, то уравнения (7.53) и система (7.55) могут быть решены только численно. Рассмотрим *метод Массо* численного решения системы (7.53), (7.55) (рис. 7.6).

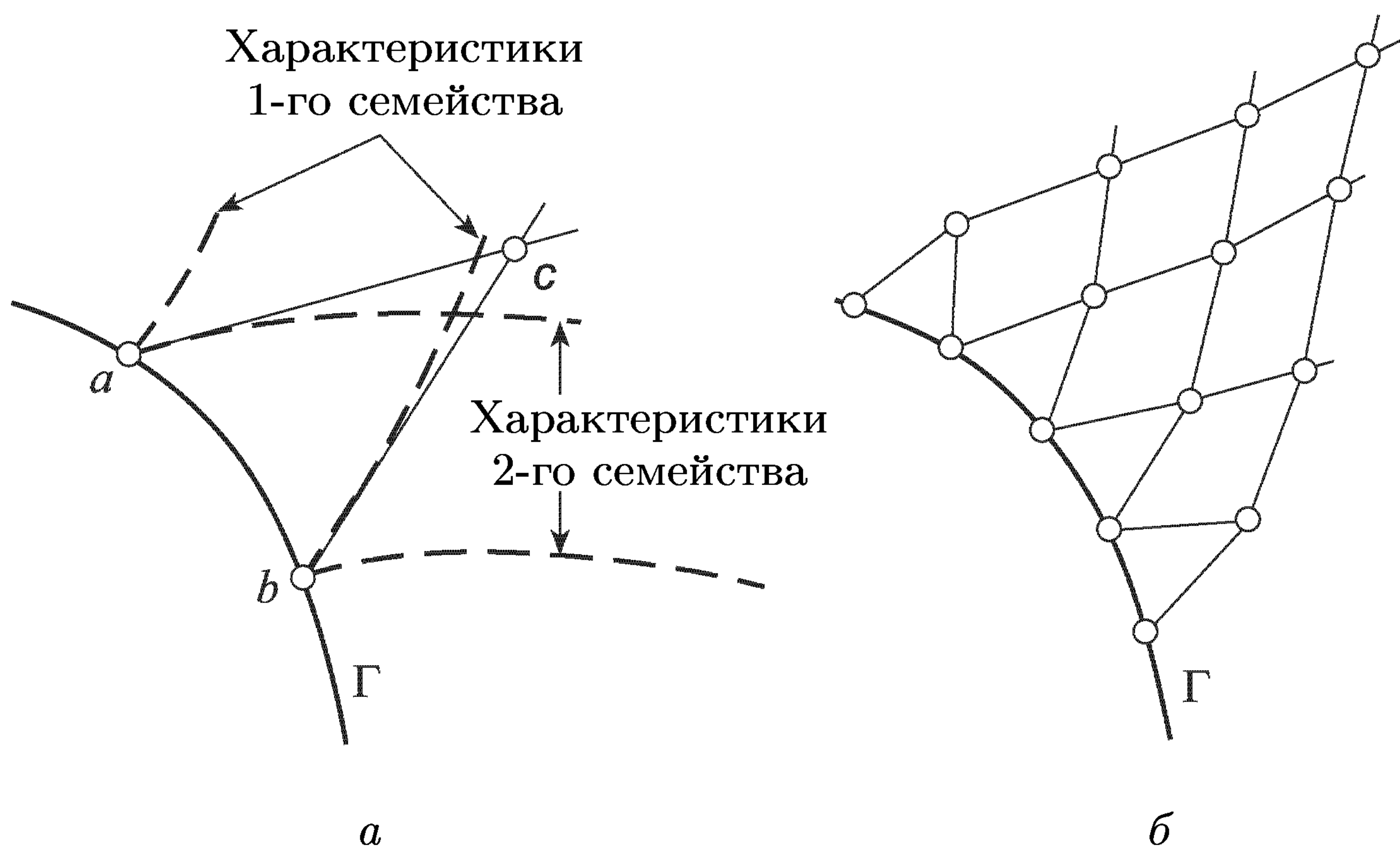


Рис. 7.6. К методу Массо

На заданной кривой Г наносится система точек с известными координатами и известными значениями функций $u\,|_\Gamma$, $v\,|_\Gamma$. Для точек $a(x_a, y_a)$ и $b(x_b, y_b)$ (рис.7.6 *а*) известны значения u_a, v_a, u_b, v_b. Из системы двух линейных алгебраических уравнений

$$\begin{cases} y_c - y_a = \lambda_{1a}(x_c - x_a), \\ y_c - y_b = \lambda_{2b}(x_c - x_b), \end{cases} \tag{7.56}$$

полученной из конечно-разностной аппроксимации двух уравнений характеристик (7.53), определяются координаты точки $c(x_c, y_c)$ как точки пересечения касательной к характеристике первого семейства, выходящей из точки a, и касательной к характеристике второго семейства, выходящей из точки b. Таким образом, точки пересечения характеристик различных семейств, выходящих из двух соседних точек, определяются как точки пересечения отрезков касательных к этим характеристикам в этих точках.

Зная координаты точки c, из конечно-разностной аппроксимации системы дифференциальных соотношений на характеристиках (7.55):

$$\begin{cases} (A\lambda_{1a} + B)(u_c - u_a) + C(v_c - v_a) + \\ \qquad + M(x_c - x_a) + N(y_c - y_a) = 0, \\ (A\lambda_{2b} + B)(u_c - u_b) + C(v_c - v_b) + \\ \qquad + M(x_c - x_b) + N(y_c - y_b) = 0, \end{cases} \tag{7.57}$$

определяются значения u_c, v_c искомых функций в точке c.

Этот алгоритм реализуется для всех точек на кривой Г. В результате получаем точки c, d, ... в окрестности кривой Г, для которых известны координаты и значения функций $u(x, y)$, $v(x, y)$. Если теперь из точки c провести отрезок касательной к характеристике первого семейства, а из соседней точки d — отрезок касательной к характеристике 2-го семейства до пересечения в некоторой новой точке, то координаты последней можно получить из решения системы (7.56), а значения функций u и v в этой точке — из решения системы (7.57). И так далее.

Таким образом, в методе характеристик отыскиваются не только значения искомых функций u и v, но и координаты точек, в которых эти решения находятся (рис.7.6 *б*).

7.7.2. Метод сквозного счета. Задача о распаде произвольного разрыва. Метод С. К. Годунова. В задачах механики сплошных сред приходится иметь дело с течениями, в которых возникают сильные разрывы (например, ударные волны, местные скачки уплотнения и т. п.), хотя начальные условия были гладкими. При расчете таких течений приходится либо

явно выделять поверхности разрыва, используя законы сохранения, либо использовать методы *сквозного счета*, позволяющие получать сквозные решения, включая поверхности разрыва.

Одним из таких методов является *метод С. К. Годунова* [24, 25], использующий *точные решения* на кусочно-постоянных начальных данных и *гибкие пространственно-временные сетки*, связанные с поверхностями разрывов.

Рассмотрим метод С. К. Годунова на примере следующей задачи Коши для уравнений акустики, являющихся следствием волнового уравнения:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{1}{\rho_0}\dfrac{\partial p}{\partial x} = 0, \quad -\infty < x < +\infty, \quad t > 0; & (7.58)\\ \dfrac{\partial p}{\partial t} + \rho_0 a_0^2 \dfrac{\partial u}{\partial x} = 0, \quad -\infty < x < +\infty, \quad t > 0; & (7.59)\\ u(x^* - 0, 0) = u_1; \quad p(x^* - 0, 0) = p_1, \quad x < x^*, \quad t = 0; & (7.60)\\ u(x^* + 0, 0) = u_2; \quad p(x^* + 0, 0) = p_2, \quad x > x^*, \quad t = 0, & (7.61) \end{cases}$$

где $u(x,t)$, $p(x,t)$ — возмущения скорости и давления в акустической волне; a_0, ρ_0 — скорость звука и плотность в невозмущенном газе; u_1, u_2, p_1, p_2 — постоянные, причем $u_1 \neq u_2$ и $p_1 \neq p_2$, т. е. в точке x^* наблюдается разрыв первого рода функций u и p.

Будем решать задачу Коши (7.58)–(7.61) методом характеристик (система (7.58), (7.59) является гиперболической системой), для чего запишем ее в форме (7.47):

$$\begin{cases} a_{11}u_t + a_{12}p_t + b_{11}u_x + b_{12}p_x = c_1,\\ a_{21}u_t + a_{22}p_t + b_{21}u_x + b_{22}p_x = c_2, \end{cases}$$

где $a_{12} = b_{11} = c_1 = a_{21} = b_{22} = c_2 = 0$; $a_{11} = a_{22} = 1$; $b_{12} = \dfrac{1}{\rho_0}$, $b_{21} = \rho_0 a_0^2$.

На начальной кривой Γ в плоскости xOt выполняются дифференциальные соотношения (7.49), из которых следуют равенства

$$u_t = \frac{du}{dt} - u_x \cdot x'(t), \tag{7.62}$$

$$p_t = \frac{dp}{dt} - p_x \cdot x'(t). \tag{7.63}$$

Подставляя (7.62), (7.63) в (7.58), (7.59), получим следующую СЛАУ относительно u_x, p_x:

$$\begin{cases} -x'u_x + \dfrac{1}{\rho_0}p_x = -\dfrac{du}{dt} & (7.64) \\[2ex] \rho_0 a_0^2 u_x - x'p_x = -\dfrac{dp}{dt} & (7.65) \end{cases}$$

$$\Delta = \begin{vmatrix} -x' & \dfrac{1}{\rho_0} \\[2ex] \rho_0 a_0^2 & -x' \end{vmatrix} = x'^2 - a_0^2 = 0, \quad x'_{1,2} = \pm a_0.$$

$x = \pm a_0 t + c$; при $t = 0$ $x = x^*$.

Таким образом, характеристиками 1-го и 2-го семейства в плоскости xOt являются прямые

$$x = \pm a_0 t + x^*. \tag{7.66}$$

Приравнивая нулю один из определителей расширенной матрицы СЛАУ (7.64), (7.65), получим два дифференциальных соотношения на характеристиках ($x'_{1,2} = \pm a_0$):

$$\Delta_1 = \begin{vmatrix} \dfrac{-du}{dt} & \dfrac{1}{\rho_0} \\[2ex] \dfrac{-dp}{dt} & \mp a_0 \end{vmatrix} = \pm a_0 \frac{du}{dt} + \frac{1}{\rho_0}\frac{dp}{dt} = 0,$$

или

$$\frac{dp}{dt} \pm a_0 \rho_0 \frac{du}{dt} = 0. \tag{7.67}$$

Проинтегрируем (7.67) по времени, получим комплексы $p \pm \pm a_0 \rho_0 u = c$ или

$$I_{\pm} = p \pm a_0 \rho_0 u, \tag{7.68}$$

называемые *инвариантами Римана*. Они сохраняют свои значения вдоль соответствующих характеристик $x = \pm a_0 t + x^*$, выходящих из точки x^* начальной кривой при $t = 0$.

Пусть в точке $(x^*, 0)$ в начальный момент $t = 0$ функции u и p претерпевают разрыв первого рода, а именно: слева от

точки $(x^*, 0)$ значения этих функций равны $u(x^* - 0, 0) = u_1$, $p(x^* - 0, 0) = p_1$, а справа $u(x^* + 0, 0) = u_2$, $p(x^* + 0, 0) = p_2$. С течением времени слева от характеристики $x + a_0 t$ (область I на рис. 7.7) значения функций $u(x,t)$ и $p(x,t)$ будут равны u_1 и p_1 соответственно, а справа от характеристики $x - a_0 t$ — значениям u_2 и p_2. Таким образом,

$$p = p_1, \qquad u = u_1, \quad x < x^* - a_0 t \quad \text{(область I)}; \tag{7.69}$$

$$p = p_2, \qquad u = u_2, \quad x > x^* + a_0 t \quad \text{(область II)}. \tag{7.70}$$

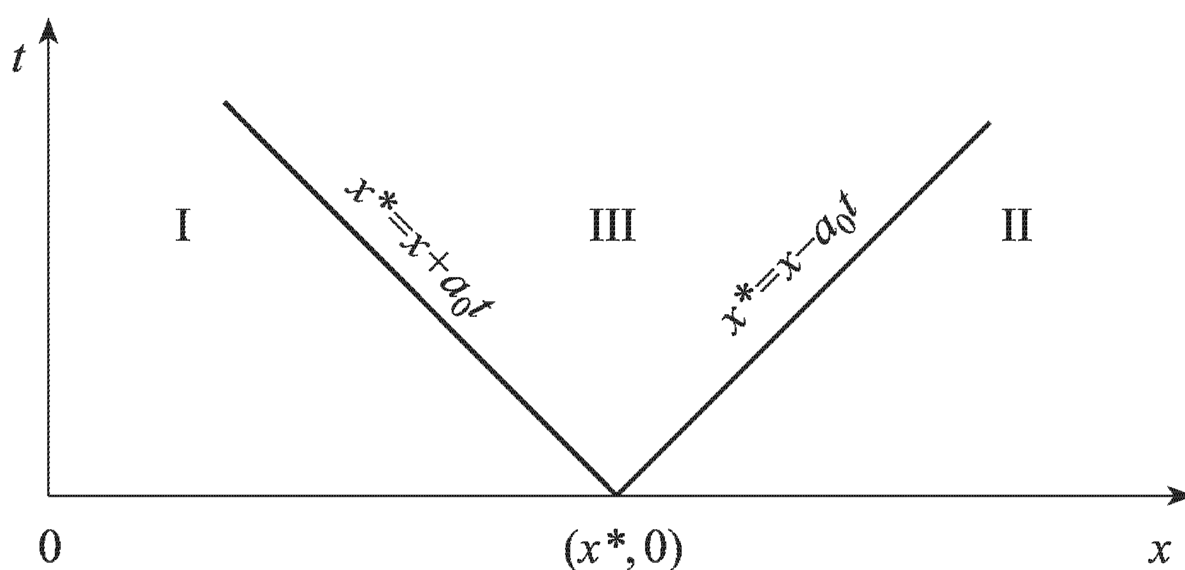


Рис. 7.7. Характеристики в точке $(x^*, 0)$

На характеристиках $x \pm a_0 t$ (рис. 7.7) в окрестности точки x^* выполняются соотношения (в соответствии с инвариантами Римана):

$$I_+ \,|_{x^*+0} = I \,|_{x^*+0} : \quad p_1 + a_0\rho_0 u_1 = p + a_0\rho_0 u; \tag{7.71}$$

$$I_- \,|_{x^*-0} = I \,|_{x^*-0} : \quad p_2 - a_0\rho_0 u_2 = p - a_0\rho_0 u. \tag{7.72}$$

Складывая (7.71), (7.72), а затем вычитая (7.72) из (7.71), получим значения функций $u(x,t)$, $p(x,t)$ внутри области III между характеристиками $x + a_0 t$ и $x - a_0 t$:

$$p = \frac{p_2 + p_1}{2} - a_0\rho_0 \frac{u_2 - u_1}{2};$$

$$u = \frac{u_2 + u_1}{2} - \frac{p_2 - p_1}{2a_0\rho_0}, \tag{7.73}$$

$$x^* - a_0 t < x < x^* + a_0 t \quad \text{(область III)}.$$

Таким образом, начальный разрыв функций $u(x,t)$, $p(x,t)$ в точке $(x^*,0)$ трансформируется в пространстве (x,t) вдоль характеристик $x^* = x \pm a_0 t$, т. е. распадается. Поэтому решения (7.69), (7.70), (7.73) задачи (7.58)–(7.61) называются *решениями задачи о распаде произвольного разрыва*. Эти решения являются обобщенными решениями задачи (7.58)–(7.61).

На основе решения задачи о распаде произвольного разрыва разработан *метод С. К. Годунова* численного решения газодинамических задач. Изложим его существо. Пусть в момент времени $t = 0$ заданы непрерывные функции $u_0(x)$, $p_0(x)$ (рис. 7.8).

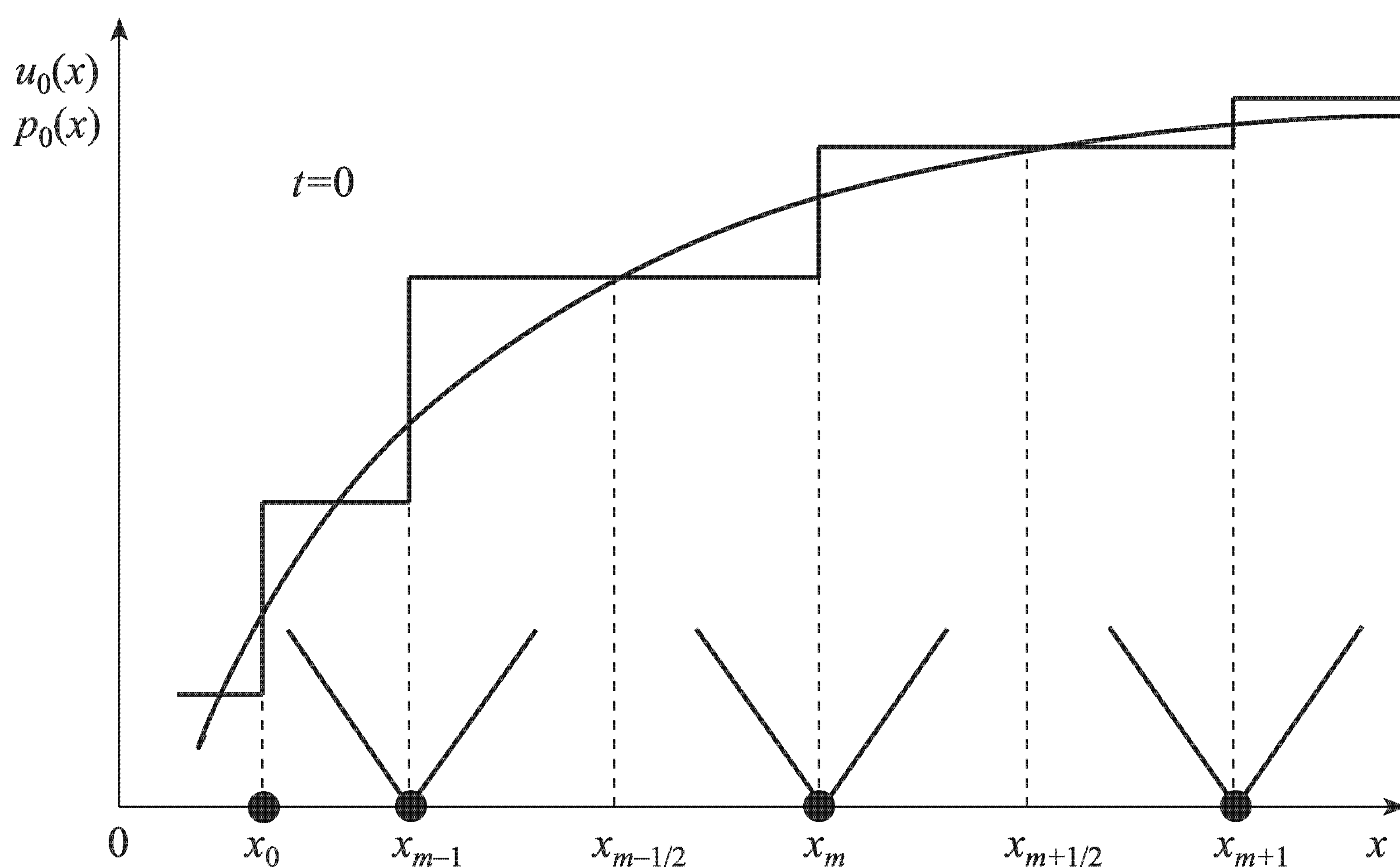


Рис. 7.8. Замена непрерывных начальных функций кусочно-постоянными

Заменим область непрерывного изменения переменной x дискретным множеством точек x_m, $m = 0, 1, 2, \ldots$, с шагами разностной сетки $h_m = x_m - x_{m-1}$, $m = 1, 2, \ldots$, и допустим, что функции $u_0(x)$, $p_0(x)$ постоянны между узлами x_{m-1}, x_m расчетной сетки и равны $p_{m-1/2}$, $u_{m-1/2}$.

В узлах x_m, $m = 0, 1, 2, \ldots$, разрыва начальных данных *возникают распады разрыва,* т. е. в каждом узле сетки, где наблюдается разрыв первого рода на начальных кривых, образуются звуковые волны, распространяющиеся вправо и влево от точек разрыва со скоростью звука a_0. Тогда, согласно решениям (7.69), (7.70), (7.73) задачи о распаде произвольного разрыва, решение

в окрестности точки x_m будет (см. рис. 7.9.) иметь вид

$$u = u_{m-1/2}, \quad p = p_{m-1/2}, \quad x_{m-1} + a_0 t < x < x_m - a_0 t; \tag{7.74}$$

$$u = u_{m+1/2}, \quad p = p_{m+1/2}, \quad x_m + a_0 t < x < x_{m+1} - a_0 t; \tag{7.75}$$

$$u = u_m = \frac{u_{m-1/2} + u_{m+1/2}}{2} - \frac{p_{m+1/2} - p_{m-1/2}}{2a_0\rho_0},$$
$$x_m - a_0 t < x < x_m + a_0 t; \tag{7.76}$$

$$p = p_m = \frac{p_{m-1/2} + p_{m+1/2}}{2} - a_0\rho_0 \frac{u_{m+1/2} - u_{m-1/2}}{2},$$
$$x_m - a_0 t < x < x_m + a_0 t. \tag{7.77}$$

На рис. 7.9. изображена структура решения (7.74)–(7.77), которая сохраняется до тех пор, пока характеристики различных

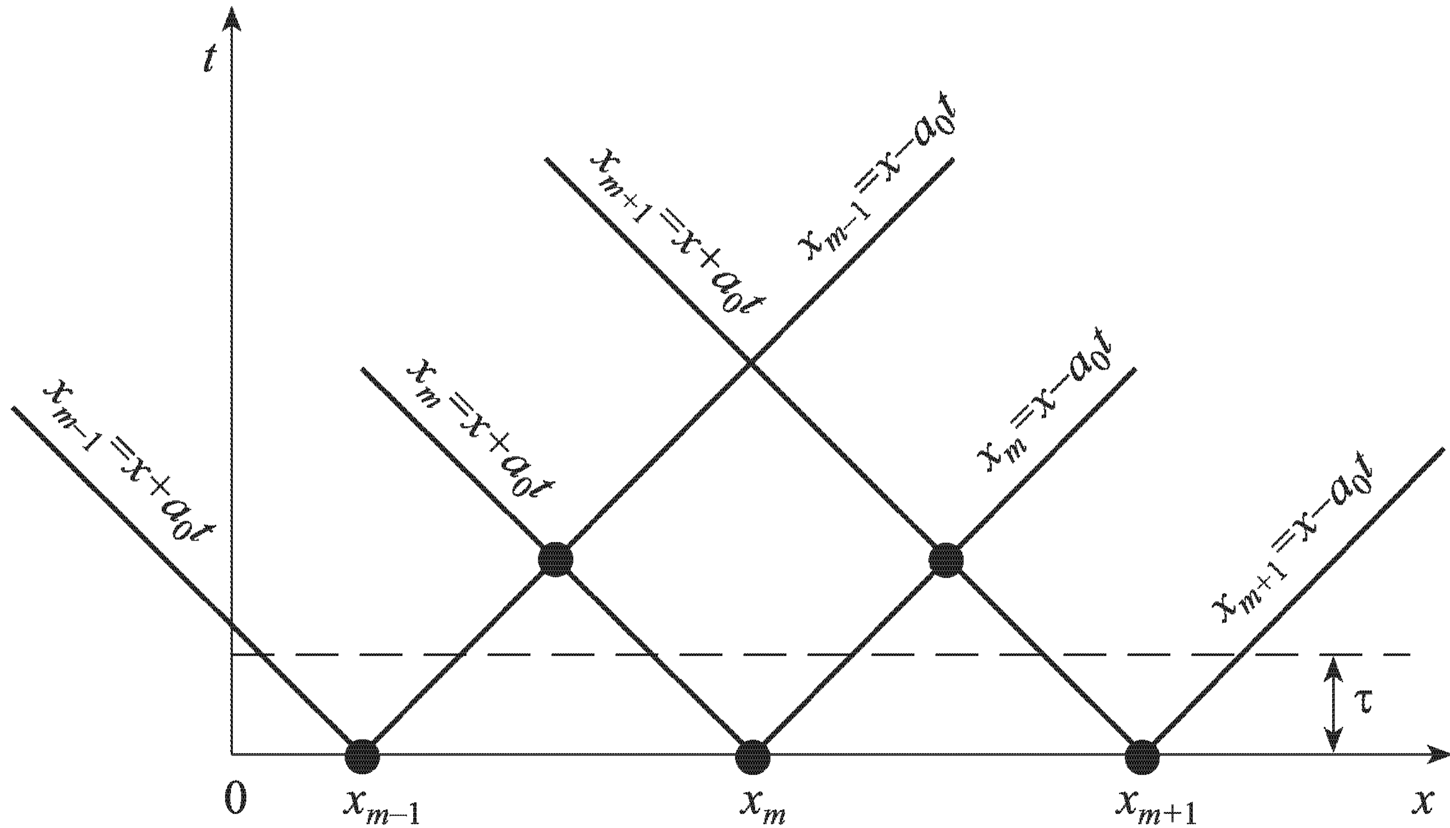


Рис. 7.9. К методу Годунова С. К.

семейств, вышедшие из соседних узлов, не пересекутся между собой (в точках пересечения получается неоднозначность решения, чего нельзя допустить).

Для предотвращения этого явления шаг по времени τ выбирается таким образом, чтобы упомянутые характеристики не пересекались. Новые средние значения $u^{m-1/2}$, $p^{m-1/2}$ на слое τ

подлежат определению с помощью интегральных законов сохранения.

Предположим, что искомые функции $u(x, t)$, $p(x, t)$, домноженные на некоторые коэффициенты, являются компонентами векторной функции на плоскости xOt. Тогда к этой векторной функции можно применить законы векторного анализа (формулу Грина на плоскости или формулы Остроградского–Гаусса и Стокса в пространстве). Применим к этому векторному полю формулу Грина, в соответствии с которой циркуляция векторного поля $\overline{a}(x, y) = P(x, y)i + Q(x, y)j$ по некоторому замкнутому контуру C, равна двойному интегралу по площади S от функции $\dfrac{\partial Q}{\partial x} - \dfrac{\partial P}{\partial y}$. Коэффициенты у компонентов векторной функции $u(x, t)$, $p(x, t)$ подбираются такими, чтобы двойные интегралы от левых частей уравнений (7.58), (7.59), равные нулю, были равны циркуляциям соответствующих векторных полей по

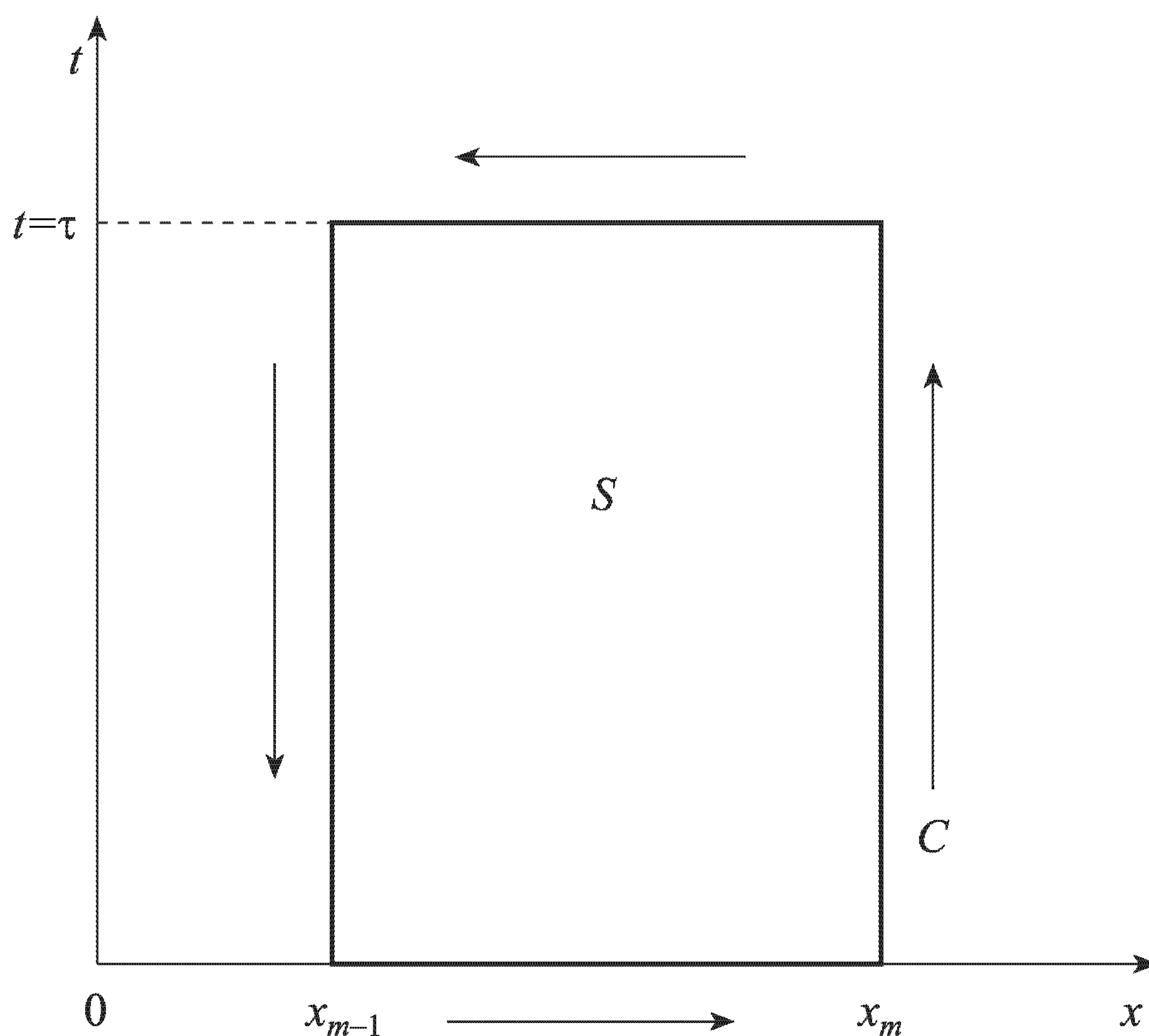


Рис. 7.10. К выбору шага τ

замкнутому контуру C. Тогда для контура, представленного на рис. 7.10, можно записать

$$\oint\limits_C \left(u dx - \frac{p}{\rho_0} dt\right) = \iint\limits_S \left(\frac{\partial u}{\partial t} + \frac{1}{\rho_0}\frac{\partial p}{\partial x}\right) dx dt = 0, \qquad (7.78)$$

$$\oint (p dx - \rho_0 a_0^2 u dt) = \iint\limits_S \left(\frac{\partial p}{\partial t} + \rho_0 a_0^2 \frac{\partial u}{\partial x} \right) dx dt = 0, \tag{7.79}$$

откуда видно, что двойные интегралы равны нулю в соответствии с уравнениями (7.58), (7.59). Записывая соотношения (7.78), (7.79) в плоскости xOt вдоль прямоугольной ячейки $x = x_{m-1}$, $x = x_m$, $t = 0$, $t = \tau$, получим для (7.78)

$$\int\limits_{x_{m-1}}^{x_m} u(x,0)dx - \int\limits_{x_{m-1}}^{x_m} u(x,\tau)dx + \\ + \int\limits_0^\tau \left(-\frac{1}{\rho_0} p(x_m, t) \right) dt - \int\limits_0^\tau \left(-\frac{p(x_{m-1}, t)}{\rho_0} \right) dt = 0,$$

или

$$u^{m-1/2} = u_{m-1/2} - \frac{\tau}{\rho_0 h}(p_m - p_{m-1}). \tag{7.80}$$

Для (7.79) циркуляция векторного поля равна

$$\int\limits_{x_{m-1}}^{x_m} p(x,0)dx - \int\limits_{x_{m-1}}^{x_m} p(x,\tau)dx - \int\limits_0^\tau \left(\rho_0 a_0^2 u(x_m, t) \right) dt + \\ + \int\limits_0^\tau \left(\rho_0 a_0^2 u(x_{m-1}, t) \right) dt = 0,$$

откуда

$$p^{m-1/2} = p_{m-1/2} - \rho_0 a_0^2 \frac{\tau}{h}(u_m - u_{m-1}). \tag{7.81}$$

В соотношениях (7.80), (7.81) верхний индекс относится к верхнему временному слою $t = \tau$, а нижний — к начальному моменту $t = 0$. Значения u_m и p_m в этих выражениях определяются соотношениями (7.76), (7.77) в задаче о распаде произвольного разрыва.

Из соотношений (7.80) и (7.81) (а также в соответствии с методом гармонического анализа исследования устойчивости) следует, что

$$\frac{\tau}{\rho_0 h} \cdot \rho_0 a_0^2 \frac{\tau}{h} \leqslant 1,$$

откуда получаем неравенство по ограничению шага численного интегрирования задачи (7.58)–(7.61) по времени:

$$\tau \leqslant \frac{h}{a_0}. \tag{7.82}$$

ГЛАВА VIII

МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ

Программа

Основы метода конечных элементов (МКЭ). Система базисных и весовых функций. Методы взвешенных невязок: коллокаций, Галеркина, наименьших квадратов. Конечно-элементный метод Галеркина решения краевых задач для ОДУ. Слабая формулировка метода Галеркина. Формирование локальной и глобальной матриц жесткости. Ансамблирование элементов. Случай граничных условий, содержащих производные. МКЭ в многомерных стационарных задачах математической физики. Принципы разбиения на конечные элементы. Ленточные матрицы жесткости. Формирование многомерных базисных функций, ансамблирование и построение глобальных СЛАУ. МКЭ в многомерных нестационарных задачах математической физики. Оценка погрешности МКЭ в задачах для ОДУ и уравнений в частных производных. Вариационный принцип в МКЭ. Вариационный принцип Релея–Ритца. Решение задач с помощью конечно-элементного вариационного принципа.

Метод конечных элементов (МКЭ) на основе вариационного принципа возник из решения задач теории упругости, что и определило, в основном, терминологию, используемую в процессе его применения в других разделах механики сплошных сред (теории теплопроводности, газовой динамике и др.) [26].

Использование в МКЭ методов взвешенных невязок (таких, например, как методы коллокаций, Галеркина, наименьших квадратов) позволило отказаться от вариационного принципа в МКЭ, тем более что не для всякой задачи можно построить функционал, минимум которого дает исследуемое дифференциальное уравнение. Тем самым круг решаемых задач механики сплошных сред был существенно расширен [27–29].

§ 8.1. Основы МКЭ

Пусть в области $\overline{\Omega} = \Omega + \Gamma$ необходимо решить некоторую дифференциальную задачу. Тогда в МКЭ осуществляется следующая цепочка процедур.

1. Область $\overline{\Omega}$ разбивают на подобласти в количестве E штук $(e = \overline{1, E})$, называемые *конечными элементами*, такие что

$$\Omega = \bigcup_{e=1}^{E} \Omega^e, \quad \Gamma = \bigcup_{e=1}^{E} \Gamma^e.$$

2. В каждом конечном элементе $\overline{\Omega}^e = \Omega^e + \Gamma^e$ выбирается *система нумерованных узлов*, в которых значения искомой функции являются неизвестными величинами.

3. Каждому нумерованному узлу приписывается *базисная функция*, такая что в этом узле она равна единице, а в остальных нумерованных узлах расчетной области — нулю. Число базисных функций в расчетной области равно числу нумерованных узлов, причем для различных узлов они обладают свойством *линейной независимости* (или *ортогональности*) по всей расчетной области.

4. Решение искомой дифференциальной задачи *приближенно* строится в виде линейной комбинации базисных функций по всем нумерованным узлам расчетной области с коэффициентами линейной комбинации, равными значениям искомой функции в нумерованных узлах.

5. Это решение подставляется в дифференциальную задачу, и, поскольку решение приближенное, результатом подстановки будет не тождественный нуль, а некоторая *функциональная невязка*.

6. С помощью известных методов *взвешенных невязок* (коллокаций, Галеркина, наименьших квадратов) функциональная невязка минимизируется по всей расчетной области путем приравнивания нулю скалярного произведения функциональной невязки и весовых функций (скалярное произведение от непрерывных функций равно определенному интегралу по расчетной области от произведения этих функций), причем в методе взвешенных невязок Галеркина весовые функции в нумерованных узлах совпадают с базисными функциями. В результате получается система линейных алгебраических уравнений (СЛАУ) относительно значений искомой функции в нумерованных узлах,

коэффициентами в которой являются интегралы по всей расчетной области от базисных функций и их производных.

7. Определенные интегралы по всей расчетной области заменяются на сумму интегралов по конечным элементам, что, в силу ортогональности базисных функций, делает матрицу СЛАУ сильно разреженной, с ненулевыми элементами, расположенными в окрестности главной диагонали (так называемые *ленточные матрицы*, частным видом которых является трехдиагональная матрица).

8. Решается СЛАУ относительно узловых значений искомой функции каким-либо известным методом (Гаусса, простых итераций, Зейделя и т. п.). Результаты решения подставляются в приближенное решение по п. 4. При этом полученные значения искомой функции в нумерованных узлах каждого конечного элемента могут быть использованы для получения решения во всех точках конечного элемента $\overline{\Omega}^e = \Omega^e + \Gamma^e$ с помощью так называемых *функций элементов*, простейшим случаем которых является линейный интерполяционный многочлен в R^n.

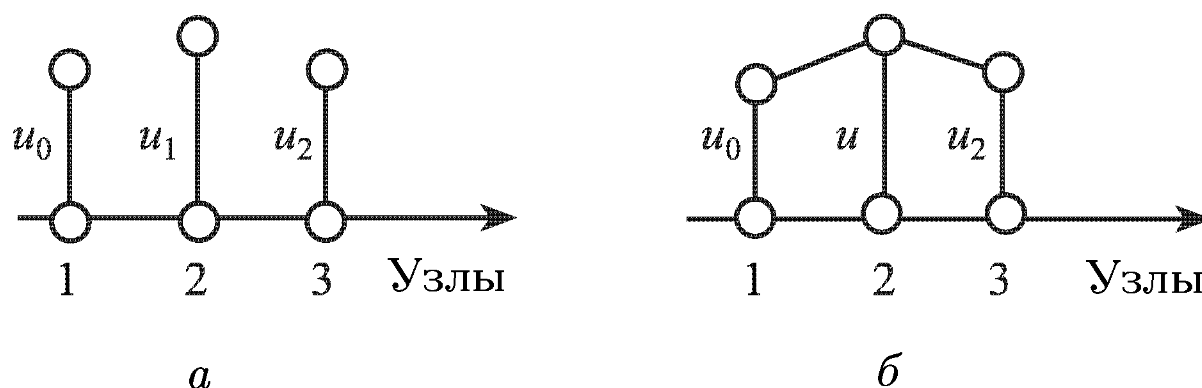


Рис. 8.1. Представление решения в МКР (*а*) и МКЭ (*б*)

Таким образом, существенным отличием МКЭ от метода конечных разностей (МКР) является то, что в МКЭ решение на каждом элементе получается в виде непрерывных (или гладких) функций, в то время как в МКР — в виде сеточной функции (рис. 8.1).

§ 8.2. Система базисных функций

В качестве базисных функций будем рассматривать два вида ортогональных базисных функций, а именно: *кусочно-постоянные базисные функции* и *линейные кусочно-непрерывные базисные функции*.

8.2.1. Кусочно-постоянные базисные функции. Пусть в вещественном пространстве R^1 рассматривается класс функций $\varphi(x)$, непрерывно дифференцируемых необходимое число раз на отрезке $x \in [0;1]$. Разобьем этот отрезок точками x_m, $m = \overline{1,M}$, на M элементарных отрезков $[x_{m-1}, x_m]$, $m = \overline{1,M}$, и представим функцию $\varphi(x)$ в виде следующей линейной комбинации:

$$\varphi(x) \approx \widehat{\varphi}(x) = \sum_{m=1}^{M} \varphi_m \cdot N_m(x), \tag{8.1}$$

где $N_m(x)$ — кусочно-постоянные функции на каждом отрезке $[x_{m-1}, x_m]$, и, если эти функции линейно-независимы (или ортогональны) при различных индексах m, будем называть их *кусочно-постоянными базисными функциями*, определяемыми равенствами

$$N_m(x) = \begin{cases} 1, \text{ если } x_{m-1} < x < x_m, \\ 0, \text{ если } x < x_{m-1} \text{ или } x > x_m, m = \overline{1,M}, \end{cases} \tag{8.2}$$

а φ_m — значения функции $\varphi(x)$ в нумерованных узлах, находящихся в середине каждого отрезка $[x_{m-1}, x_m]$.

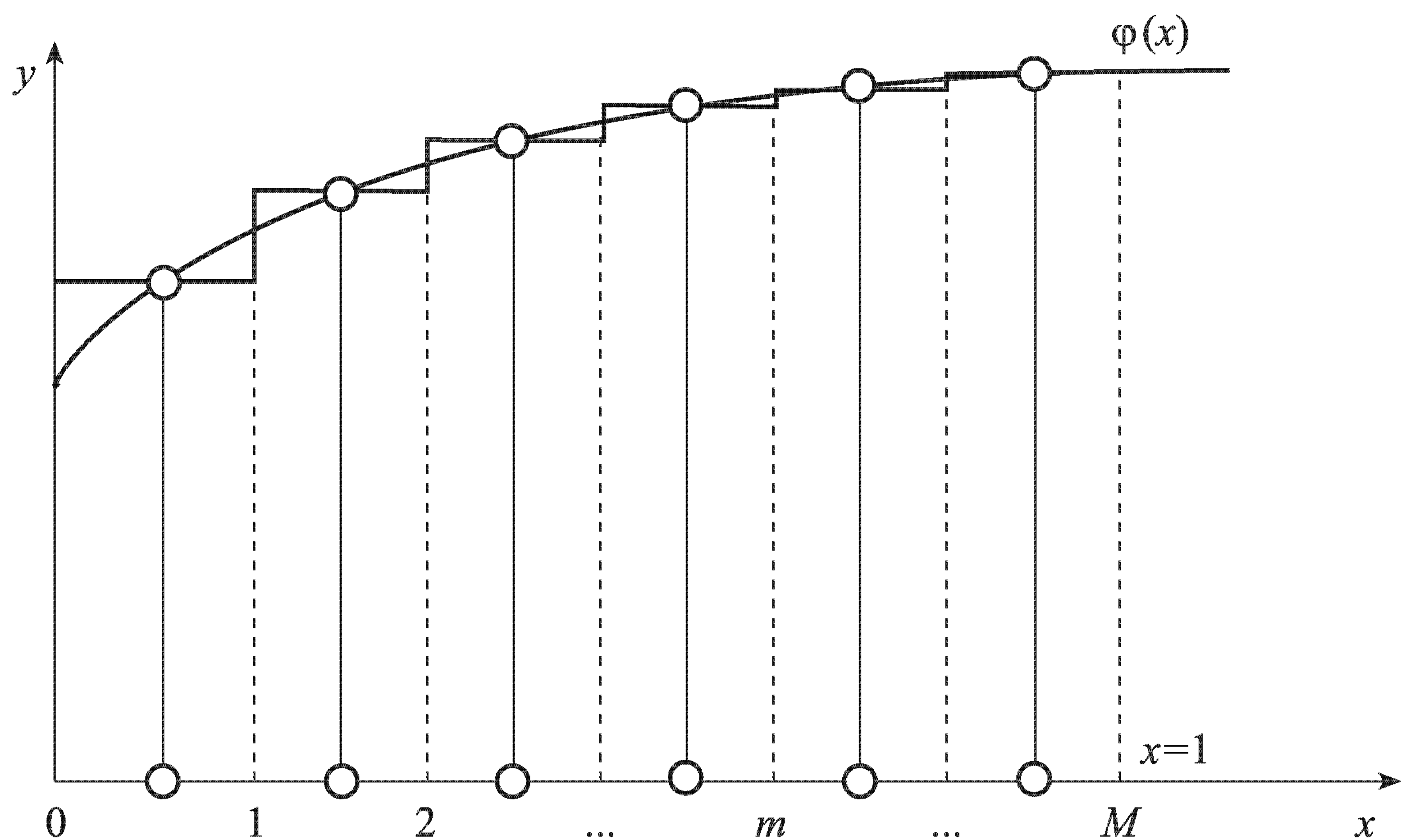


Рис. 8.2. Аппроксимация функции с помощью кусочно-постоянных базисных функций

Тогда аппроксимация (8.1) функции $\varphi(x)$ на отрезке $x \in [0;1]$ с помощью кусочно-постоянных базисных функций (8.2) геометрически представляется ступенчатой функцией (рис. 8.2). При

этом каждая базисная функция, приписанная нумерованному узлу, принимает значение, равное единице только на отрезке, внутри которого расположен этот нумерованный узел; во всех остальных нумерованных узлах расчетной области эта базисная

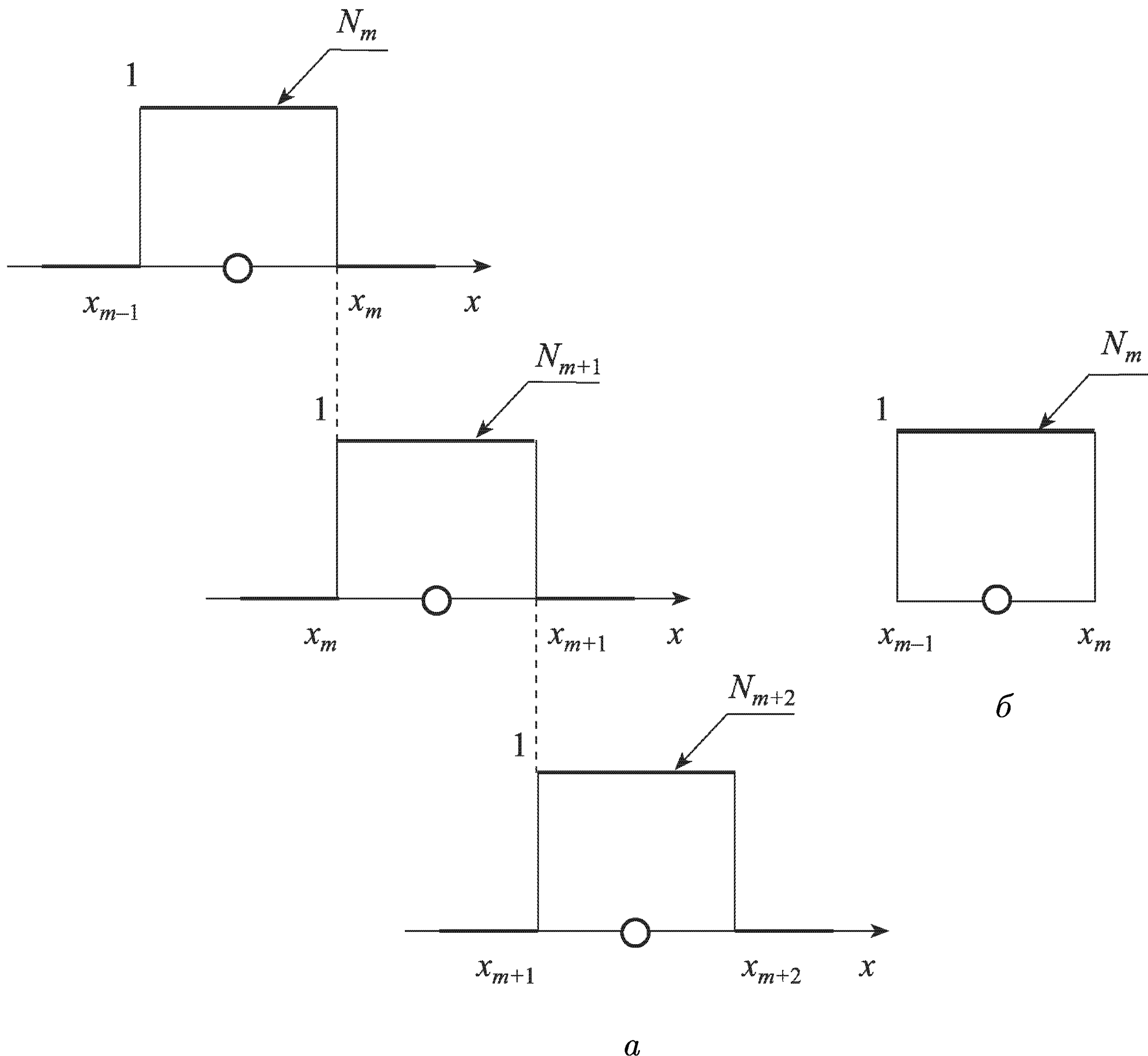


Рис. 8.3. Глобальные (*а*) и локальные (*б*) кусочно-постоянные базисные функции

функция равна нулю (рис. 8.3). Такие базисные функции называют *глобальными* (рис. 8.3 *а*), т. е. заданными в форме (8.2) на всей расчетной области.

В отличие от этого, базисную функцию, заданную на отрезке $[x_{m-1}, x_m]$ (рис. 8.3 *б*), называют *локальной* базисной функцией.

Как видно из определения (8.2) глобальных кусочно-постоянных базисных функций и из рис. 8.3 *а*, эти функции ортогональны на отрезке $x \in [0; 1]$ в смысле скалярного произведения, т. е. для двух кусочно-постоянных базисных функций

с номерами i и j $(i \neq j,\ i, j = \overline{1, M})$ имеет место равенство

$$(N_i(x), N_j(x)) = \int\limits_0^1 N_i(x)N_j(x)dx = \int\limits_0^{x_{i-1}} N_i(x)N_j(x)dx +$$

$$+ \int\limits_{x_{i-1}}^{x_i} N_i(x)N_j(x)dx + \int\limits_{x_i}^{x_{j-1}} N_i(x)N_j(x)dx + \int\limits_{x_{j-1}}^{x_j} N_i(x)N_j(x)dx +$$

$$+ \int\limits_{x_j}^{1} N_i(x)N_j(x)dx = 0, \quad i, j = \overline{1, M}, \quad i \neq j.$$

8.2.2. Линейные кусочно-непрерывные базисные функции. Если в качестве базисных функций принять функции вида

$$N_m(x) = \begin{cases} \dfrac{x - x_{m-1}}{x_m - x_{m-1}}, x \in [x_{m-1}, x_m]\,; \\[2ex] \dfrac{x_{m+1} - x}{x_{m+1} - x_m}, x \in [x_m, x_{m+1}]\,; \\[2ex] 0, x < x_{m-1}, x > x_{m+1}, \end{cases} \tag{8.3}$$

называемые *линейными кусочно-непрерывными базисными функциями*, то аппроксимация (8.1) функции $\varphi(x)$ на каждом отрезке $x \in [x_{m-1}, x_m]$ будет линейной и непрерывной в узлах x_m. В этом случае в качестве нумерованных узлов принимаются узлы разбиения (рис. 8.4).

Для отдельного отрезка $x \in [x_m, x_{m+1}]$ с нумерованными узлами m и $m + 1$ глобальные линейные кусочно-непрерывные базисные функции представлены на рис. 8.5 *а*, а локальные — на рис. 8.5 *б*. Из рис. 8.5 видно, что глобальные базисные функции, построенные для различных узлов и действующие на различных отрезках, ортогональны на всей области изменения переменной x.

Можно построить и другие базисные функции, в том числе и нелинейные.

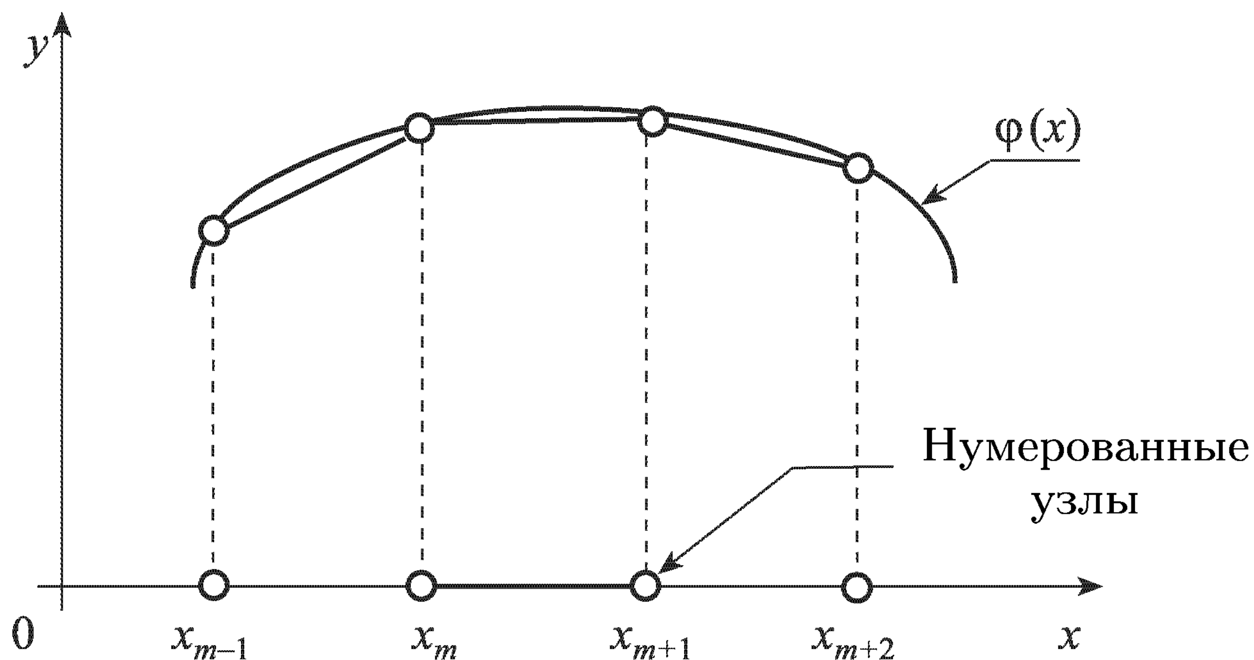


Рис. 8.4. Аппроксимация функции с использованием линейных кусочно-непрерывных функций

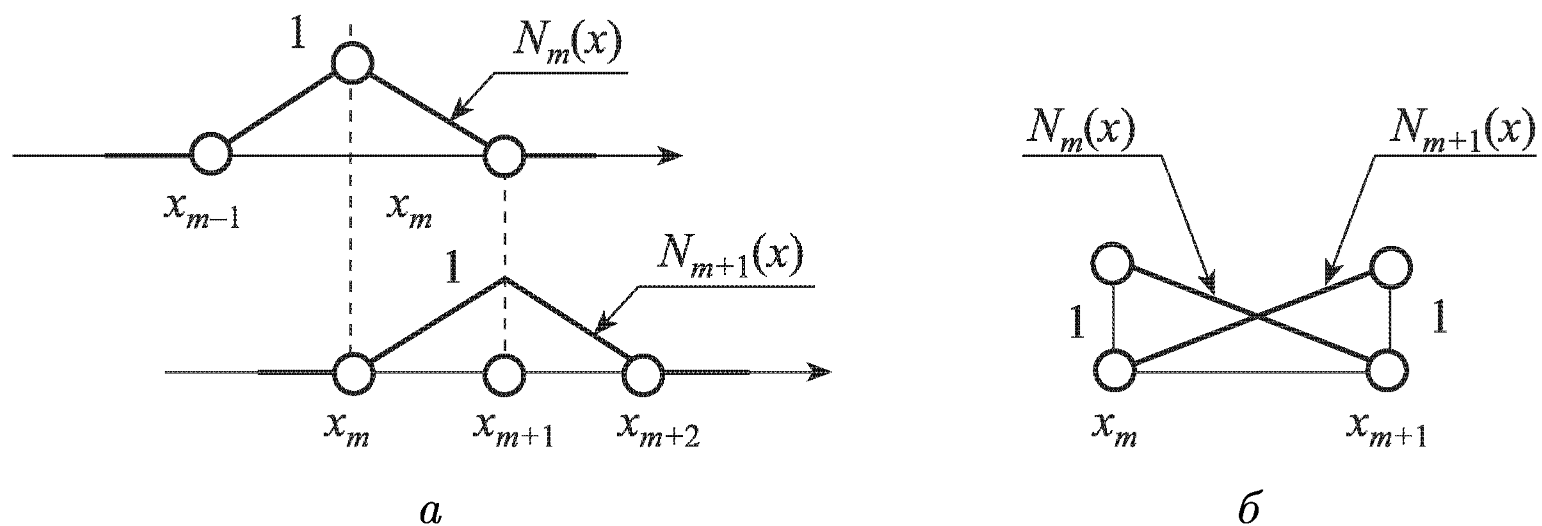


Рис. 8.5. Глобальные (*а*) и локальные (*б*) линейные кусочно-непрерывные базисные функции

§ 8.3. Методы взвешенных невязок. Весовые функции

Пусть в области $\overline{\Omega} = \Omega + \Gamma \subset R^n$ рассматривается краевая задача

$$Lu + p = 0 \text{ в } \Omega; \tag{8.4}$$

$$Bu + q = 0 \text{ на } \Gamma, \tag{8.5}$$

где L и B — линейные дифференциальные операторы, а p и q — известные функции независимых переменных.

Для аппроксимации функции u в нумерованных узлах области Ω выбирается система базисных функций N_m, $m = \overline{1, M}$, с помощью которой строится приближенное решение

$$u \approx \widehat{u} = \sum_{m=1}^{M} u_m \cdot N_m, \tag{8.6}$$

где u_m — значения искомой функции в нумерованных узлах.

Подставляя это приближенное решение в задачу (8.4), (8.5), получим не тождественный нуль, поскольку (8.6) — приближенное решение, а некоторую функциональную невязку R_Ω по области Ω и невязку R_Γ по границе Γ:

$$R_\Omega = \sum_{m=1}^{M} u_m L(N_m) + p; \quad R_\Gamma = \sum_{m=1}^{M} u_m B(N_m) + q.$$

С целью нахождения искомых значений u_m, $m = \overline{1, M}$, ортогонализуем невязки R_Ω и R_Γ с помощью специальным образом подобранных функций W_s, $s = \overline{1, M}$, называемых *весовыми функциями*, т. е. приравниваем нулю скалярные произведения невязок R_Ω и R_Γ и весовых функций W_s, $s = \overline{1, M}$, в результате чего получается следующая система линейных алгебраических уравнений относительно узловых значений u_m искомой функции:

$$\begin{aligned}(W_s, R_{\overline{\Omega}}) = \int_\Omega W_S \cdot R_\Omega d\Omega + \int_\Gamma \overline{W}_s \cdot R_\Gamma d\Gamma = \\ = \int_\Omega W_s \left(\sum_{m=1}^{M} u_m L\left(N_m\right) + p\right) d\Omega + \\ + \int_\Gamma \overline{W}_s \cdot \left(\sum_{m=1}^{M} u_m B\left(N_m\right) + q\right) d\Gamma = 0, \quad s = \overline{1, M},\end{aligned} \tag{8.7}$$

где весовые функции W_S и $\overline{W}_S$ для области Ω и границы Γ могут быть разными.

Этот метод ортогонализации невязок R_Γ и R_Ω с помощью весовых функций называют *методом взвешенных невязок*, различные варианты которого отличаются способом задания весовых функций W_s, $s = \overline{1, M}$.

8.3.1. Метод поточечной коллокации. В расчетной области $\overline{\Omega}$ выбирается $s = \overline{1, M}$ *точек коллокации* (совпадающих с нумерованными узлами расчетной области), в которых невязка $R_{\overline{\Omega}}$ полагается равной нулю. В этом случае в качестве весовых

функций принимаются дельта-функции Дирака, которые в n-мерном вещественном пространстве имеют вид

$$W_s = \delta(r - r_s), s = \overline{1, M},$$

где r_s — длина радиуса-вектора точки коллокации, а дельта-функция $\delta(r - r_s)$ обладает следующими свойствами:

$$\delta(r - r_s) = \begin{cases} 0, & \text{если } r \neq r_s; \\ \infty, & \text{если } r = r_s; \end{cases}$$

$$\int\limits_{r<r_s}^{r>r_s} \delta(r - r_s) dr = 1, \quad s = \overline{1, M}; \tag{8.8}$$

$$\int\limits_{r<r_s}^{r>r_s} R_{\overline{\Omega}}(r)\delta(r - r_s) dr = R_{\overline{\Omega}}(r_s), \quad s = \overline{1, M}. \tag{8.9}$$

В соответствии со свойством (8.9) и условием ортогональности (8.7) получаем следующую систему из M линейных алгебраических уравнений относительно узловых значений u_m искомой функции:

$$\int\limits_{\overline{\Omega}} R_{\overline{\Omega}}(r)\delta(r - r_s) d\overline{\Omega} = R_{\overline{\Omega}}(r_s) = 0, \quad s = \overline{1, M}.$$

8.3.2. Метод Галеркина. В этом методе в качестве весовых функций выбираются базисные функции в нумерованных узлах расчетной области $W_s = N_s$, $s = \overline{1, M}$.

В результате СЛАУ (8.7) относительно u_m приобретает вид

$$\int\limits_{\Omega} N_s \left(\sum_{m=1}^{M} u_m L(N_m) + p \right) d\Omega + \\ + \int\limits_{\Gamma} \overline{N}_s \cdot \left(\sum_{m=1}^{M} u_m B(N_m) + q \right) d\Gamma = 0, \quad s = \overline{1, M},$$

причем, в силу ортогональности базисных функций, матрица этой СЛАУ будет иметь разреженный вид.

Среди методов взвешенных невязок метод Галеркина является одним из наиболее популярных.

8.3.3. Метод наименьших квадратов. Метод наименьших квадратов заключается в минимизации функционала, являющегося интегралом по области $\overline{\Omega}$ от квадрата невязки $R_{\overline{\Omega}}$:

$$I(u_1, u_2, \dots, u_M) = \frac{1}{2}\int\limits_{\overline{\Omega}} R_{\overline{\Omega}}^2 d\overline{\Omega}.$$

Необходимым условием минимума этого функционала является равенство нулю частных производных 1-го порядка по параметрам $u_1, u_2, \dots, u_M$, а учитывая, что $\dfrac{\partial \widehat{u}}{\partial u_s} = N_s$, $s = \overline{1, M}$, приходим к следующей СЛАУ относительно u_m, $m = \overline{1, M}$:

$$\int\limits_{\Omega} R_{\Omega} \cdot L\left(N_s\right) d\Omega + \int\limits_{\Gamma} R_{\Gamma} \cdot B\left(\overline{N_s}\right) d\Gamma = 0, \quad s = \overline{1, M}.$$

§ 8.4. Конечно-элементный метод Галеркина решения краевых задач для обыкновенных дифференциальных уравнений

Все содержание метода конечных элементов ниже излагается на примере следующей первой краевой задачи для обыкновенного дифференциального уравнения (ОДУ):

$$\frac{d^2u}{dx^2} - u = 0, \quad 0 < x < 1; \tag{8.10}$$

$$u(0) = 0, \quad x = 0, \tag{8.11}$$

$$u(1) = 1, \quad x = 1, \tag{8.12}$$

допускающей аналитическое решение, которое имеет вид

$$u(x) = [\exp(1 - x) - \exp(1 + x)]/(1 - e^2).$$

Разобьем отрезок $[0; 1]$ на три конечных элемента ($e = \overline{1, E}$; $E = 3$), при этом нумерованные узлы m принимаются в точках разбиения, т. е. m принимает значения 1, 2, 3, 4.

Каждому нумерованному узлу приписывается базисная функция $N_m(x)$, $m=\overline{1,4}$, определяемая с помощью выражения (8.3). Решение задачи (8.10)–(8.12) записывается в виде

$$u\,(x)\approx\widehat{u}\,(x)=\sum_{m=1}^{4}u_m\cdot N_m(x), \tag{8.13}$$

где u_m — значения искомой функции в нумерованных узлах, подлежащие определению.

Если они определены, то на конечных элементах с номерами $e=\overline{1,E}$ и произвольными нумерованными узлами m и $m+1$ искомая функция в точках этих элементов описывается следующими линейными функциями (можно показать, что они совпадают с интерполяционными многочленами первой степени):

$$u^e(x)=u_m^e N_m^e(x)+u_{m+1}^e N_{m+1}^e(x),\quad e=1,2,3.$$

Такие функции называют *функциями элемента.*

Подставляя (8.13) в ОДУ (8.10) получаем следующую функциональную невязку:

$$R_\Omega=\frac{d^2\widehat{u}}{dx^2}-\widehat{u}.$$

Поскольку на границах заданы граничные условия первого рода, выполняющиеся точно, то невязка R_Γ на границах равна нулю.

В соответствии с методом взвешенных невязок Галеркина весовые функции $W_s(x)$, $s=\overline{1,4}$, принимаются равными базисным функциям $N_s(x)$, $s=\overline{1,4}$ ($W_s(x)=N_s(x)$). Тогда в соответствии с методом Галеркина получаем следующую СЛАУ относительно u_m, $m=\overline{1,4}$:

$$\int_0^1 N_s\left(\frac{d^2\widehat{u}}{dx^2}-\widehat{u}\right)dx=0,\,s=1,2,3,4. \tag{8.14}$$

8.4.1. Слабая формулировка метода Галеркина. Известно, что определенный интеграл существует в случаях, когда подынтегральная функция является кусочно-непрерывной на заданном отрезке с конечным числом точек разрыва первого рода и ограниченной (разрывы второго рода в подынтегральных функциях не допустимы).

Однако во вторую производную подынтегральных функций СЛАУ (8.14) входят линейные кусочно-непрерывные базисные

функции $N_m(x)$, $m = \overline{1,4}$ (8.3), производные второго порядка от которых претерпевают разрывы второго рода (стремятся к $\pm\infty$). Возникает вопрос, допустимо ли использование базисных функций вида (8.3) в аппроксимации (8.13), т. е. достаточна ли гладкость базисных функций в классе C^0 (в классе непрерывных функций)?

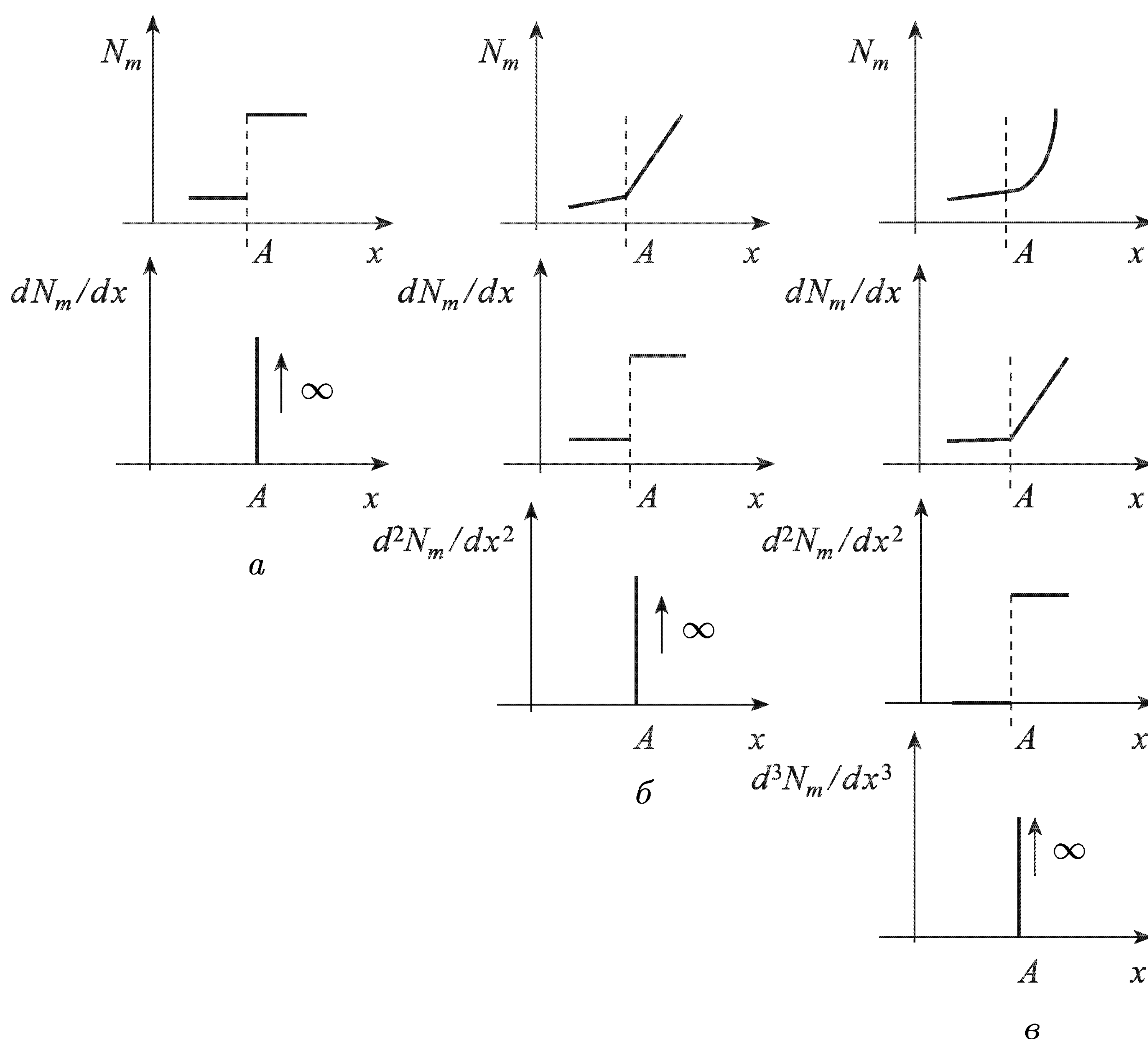


Рис. 8.6. Поведение функций и их производных: *а* — разрывных, *б* — непрерывных, *в* — непрерывно дифференцируемых

Для ответа на этот вопрос рассмотрим поведение трех типов функций $N_m(x)$ и их производных до третьего порядка включительно в окрестности точки A стыковки двух соседних элементов (рис. 8.6), а именно разрывных, непрерывных (класса C^0) и непрерывно дифференцируемых (класса C^1).

Из рис. 8.6 видно, что класс разрывных функций (рис. 8.6 *а*) не годится для аппроксимации дифференциальной задачи (8.10)–(8.12), поскольку уже первая производная претерпевает

разрыв второго рода. То же самое относится и к классу непрерывных функций (рис. 8.6 *б*) поскольку вторые производные претерпевают разрыв второго рода. Класс непрерывно дифференцируемых функций (рис. 8.6 *в*) можно использовать для конечно-элементной аппроксимации дифференциальной задачи (8.10)–(8.12), поскольку производные второго порядка являются кусочно-непрерывными с разрывами первого рода, т. е. определенные интегралы от вторых производных таких функций существуют.

Таким образом, в МКЭ на гладкость базисных функций накладываются жесткие ограничения. Покажем, как можно ослабить требования к гладкости базисных функций. Такие способы ослабления гладкости базисных функций называют «слабой формулировкой» в МКЭ. Эти способы в одномерных задачах основаны на интегрировании по частям, а в многомерных — на использовании формул Грина.

Проинтегрируем в (8.14) первое слагаемое по частям, получим

$$-\int_0^1 \left(\frac{dN_s}{dx}\cdot\frac{d\widehat{u}}{dx} + N_s\widehat{u}\right)dx + \left[N_s\frac{d\widehat{u}}{dx}\right]_0^1 = 0, \quad s = \overline{1, M+1}, \tag{8.15}$$

откуда видно, что теперь требуется непрерывность только приближенного решения $\widehat{u}$ (а следовательно, и базисных функций N_m, входящих в $\widehat{u}$ в соответствии с (8.13)), и его первой производной $\dfrac{d\,\widehat{u}}{d\,x}$.

В СЛАУ (8.15) в сумме для $\widehat{u}$ при каждом $s = \overline{1, M+1}$, перебираются все $m = \overline{1, M+1}$, где s — номер уравнения (и одновременно номер весовой функции), а m — номер неизвестного u_m в (8.13) (и одновременно номер базисной функции).

8.4.2. Формирование локальной и глобальной матриц жесткости. Ансамблирование элементов. Представим СЛАУ (8.15) в следующей векторно-матричной форме:

$$Ku = f, \;\; u = (u_1\, u_2\, \ldots\, u_{M+1})^T, \;\; f = (f_1\, f_2\, \ldots\, f_{M+1})^T, \tag{8.16}$$

где $K = [k_{sm}]$, s, $m = \overline{1, M+1}$, причем элементы матрицы K и компоненты вектора правых частей вычисляются следующим образом (достаточно в (8.15) подставить (8.13)):

$$k_{sm} = \int\limits_0^1 \left(\frac{dN_s}{dx} \frac{dN_m}{dx} + N_s N_m \right) dx, \quad s,\, m = \overline{1, M+1};$$

$$f_s = \left[N_s \frac{d\widehat{u}}{dx} \right]_0^1 = N_s \left. \frac{d\widehat{u}}{dx} \right|_{x=1} - N_s \left. \frac{d\widehat{u}}{dx} \right|_{x=0}, \quad s = \overline{1, M+1}.$$

Элементы k_{sm} и компоненты f_s вычисляются таким образом, что для s-й весовой функции N_s, $s = \overline{1, M+1}$, перебираются все базисные функции N_m, $m = \overline{1, M+1}$, и после интегрирования образуется s-я строка матрицы K и s-й элемент в векторе f.

В каждый коэффициент k_{sm} и компонент f_s в общем случае делает вклад каждый конечный элемент расчетной области с использованием выражений

$$k_{sm} = \sum_{e=1}^{E} k_{sm}^e, \quad f_s = \sum_{e=1}^{E} f_s^e,$$

где

$$k_{sm}^e = k_{ms}^e = \int\limits_{x_m}^{x_{m+1}} \left(\frac{dN_s^e}{dx} \frac{dN_m^e}{dx} + N_s^e N_m^e \right) dx, \quad s \neq m, \tag{8.17}$$

$$k_{mm}^e = \int\limits_{x_m}^{x_{m+1}} \left(\left(\frac{dN_m^e}{dx} \right)^2 + (N_m^e)^2 \right) dx, \quad s = m. \tag{8.18}$$

Для конечного элемента с нумерованными узлами m и $m+1$ построим с помощью (8.17), (8.18) матрицу K^e:

$$K^e = \int\limits_{x_m}^{x_{m+1}} \begin{bmatrix} \left(\dfrac{dN_m^e}{dx} \right)^2 + (N_m^e)^2 & \boxed{\begin{array}{c} \dfrac{dN_m^e}{dx} \cdot \dfrac{dN_{m+1}^e}{dx} + \\ + N_m^e \cdot N_{m+1}^e \end{array}} \\ \boxed{\begin{array}{c} \dfrac{dN_{m+1}^e}{dx} \cdot \dfrac{dN_m^e}{dx} + \\ + N_{m+1}^e \cdot N_m^e \end{array}} & \left(\dfrac{dN_{m+1}^e}{dx} \right)^2 + \left(N_{m+1}^e \right)^2 \end{bmatrix} dx, \tag{8.19}$$

причем для $e = 1$ $m = 1$, $e = 2$ $m = 2$, $e = 3$ $m = 3$.

Матрицу (8.19) называют *матрицей элемента* или *локальной матрицей жесткости*, в отличие от матрицы K в (8.16), называемой *глобальной матрицей жесткости*, построенной для всей расчетной области.

Процесс суммирования локальных матриц жесткости, называемый *ансамблированием конечных элементов*, осуществляется таким образом, чтобы каждая поэлементная составляющая коэффициента k_{sm} глобальной матрицы жесткости K была на своем месте. Рассмотрим процесс ансамблирования в задаче (8.10)–(8.12), для которой $E = M = 3$ (рис. 8.7).

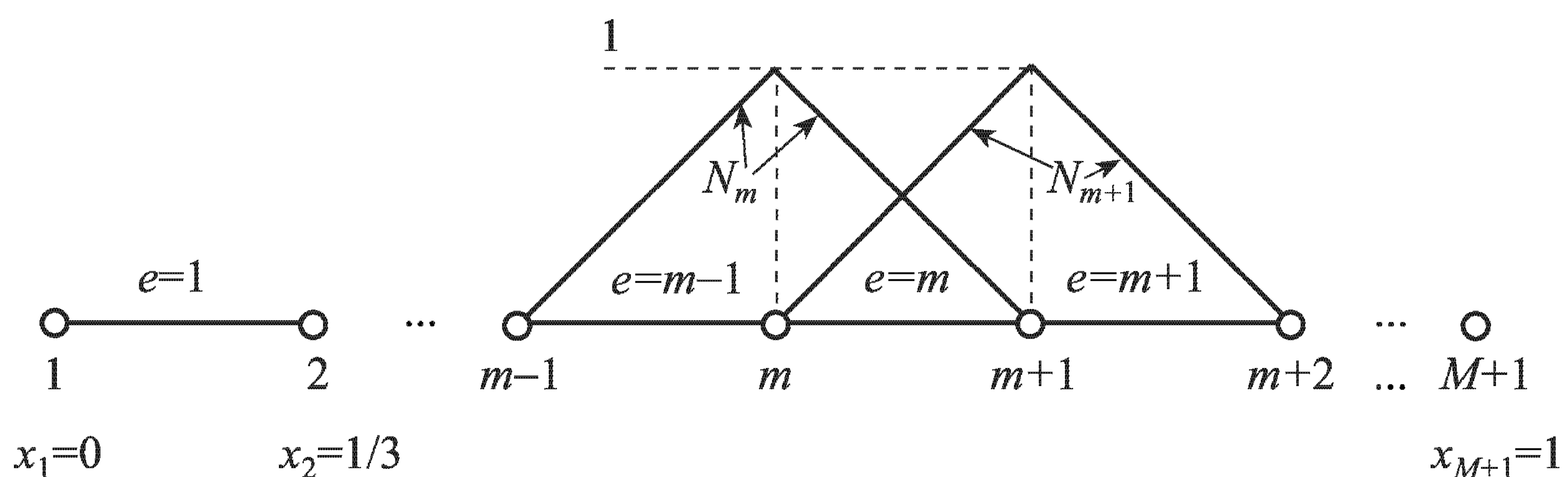


Рис. 8.7. К вычислению локальных матриц жесткости

1) Пусть $e = 1$, тогда при $s = 1$ $m = 1$, $m + 1 = 2$; при $s = 2$ $m = 1$, $m + 1 = 2$.

Локальная матрица жесткости (8.19) в узлах 1-го конечного элемента вычисляется следующим образом. Элементы матрицы (8.19), стоящие в s-й строке и m-м столбце (s, $m = 1, 2$) вычисляются по правой ветви базисной функции (8.3), приписанной левому нумерованному узлу, и по левой ветви базисной функции (8.3), приписанной правому нумерованному узлу (на рис. 8.7 эти ветви для m-го и $(m + 1)$-го узлов нанесены жирно).

Таким образом, элементы этой матрицы равны ($x_{m+1} - x_m = 1/3$, $m = 1, 2, 3$)

$$K^1 = \int\limits_0^{1/3} \begin{pmatrix} (-3)^2 + \left(\dfrac{x_2 - x}{1/3}\right)^2 & \boxed{\begin{matrix} (-3)\cdot 3 + \\ + \left(\dfrac{x_2 - x}{1/3}\right)\left(\dfrac{x - x_1}{1/3}\right) \end{matrix}} \\ \boxed{\begin{matrix} 3\cdot(-3) + \\ + \left(\dfrac{x - x_1}{1/3}\right)\left(\dfrac{x_2 - x}{1/3}\right) \end{matrix}} & 3^2 + \left(\dfrac{x - x_1}{1/3}\right)^2 \end{pmatrix} dx =$$

$$= \begin{bmatrix} 28/9 & -53/18 \\ -53/18 & 28/9 \end{bmatrix}.$$

Вектор правых частей глобальной СЛАУ содержит второе слагаемое выражения (8.15), имеющего место только для граничных узлов:

$$\left[N_s \frac{d\widehat{u}}{dx}\right]_0^1 = N_4 \frac{d\widehat{u}}{dx}\bigg|_{x=1} - N_1 \frac{d\widehat{u}}{dx}\bigg|_{x=0}.$$

Первый элемент содержит только левую границу, и, следовательно,

$$f^1 = \left(f_1^1\ f_2^1\right)^T = \left(-N_1 \frac{d\widehat{u}}{dx}\bigg|_{x=0}\ \ 0\right)^T.$$

Таким образом, локальная СЛАУ для первого конечного элемента имеет вид

$$\begin{bmatrix} 28/9 & -53/18 \\ -53/18 & 28/9 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -\left(d\widehat{u}/dx\right)\big|_{x=0} \\ 0 \end{bmatrix}. \qquad (8.20)$$

Эта локальная СЛАУ, построенная для элемента $e = 1$, учитывается в глобальной СЛАУ (8.16) следующим образом:

$$\begin{matrix} & m=1 & m=2 & & \\ s=1 & & & & \\ s=2 & & & & \end{matrix} \begin{bmatrix} 28/9 & -53/18 & 0 & 0 \\ -53/18 & 28/9 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} -\left(d\widehat{u}/dx\right)\big|_{x=0} \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (8.21)$$

т. е. элементы матрицы K^1 и вектора правых частей f^1 расширяются до размеров глобальной матрицы $K_{(M+1)\times(M+1)}$ и глобального вектора f_{M+1}, где $M = 3$, причем при $s = 1$ ($\mathrm{m} = 1$, $\mathrm{m} + 1 = 2$) первая строка матрицы K^1 ставится в первую строку и в 1-й и 2-й столбцы расширенной матрицы. При $s = 2$ ($m = 1$, $m + 1 = 2$) соответствующие элементы ставятся во вторую строку расширенной матрицы.

2) Пусть $e = 2$, тогда при $s = 2$ $m = 2$, $m + 1 = 3$; при $s = 3$ $m = 2$, $m + 1 = 3$.

Аналогично элементу $e = 1$ из (8.19) имеем

$$K^2 = \begin{bmatrix} 28/9 & -53/18 \\ -53/18 & 28/9 \end{bmatrix}; \quad f^2 = (0\ 0)^T,$$

$$\begin{array}{r} \\ \\ s=2 \\ s=3 \\ \\ \end{array}\begin{array}{c} \qquad\quad m=2 \quad\; m=3 \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 28/9 & -53/18 & 0 \\ 0 & -53/18 & 28/9 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \qquad (8.22)$$

3) Для $e = 3$ при $s = 3$ $m = 3$, $m + 1 = 4$; при $s = 4$ $m = 3$, $m + 1 = 4$, из (8.19) получаем

$$K^3 = \begin{bmatrix} 28/9 & -53/18 \\ -53/18 & 28/9 \end{bmatrix}; \quad f^3 = \left(0\ (d\widehat{u}/dx)\big|_{x=1}\right)^T,$$

$$\begin{array}{r} \\ \\ \\ s=3 \\ s=4 \end{array}\begin{array}{c} \qquad\qquad m=3 \quad\; m=4 \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 28/9 & -53/18 \\ 0 & 0 & -53/18 & 28/9 \end{bmatrix} \end{array} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ (d\widehat{u}/dx)_{x=1} \end{bmatrix}. \qquad (8.23)$$

Завершается процесс ансамблирования обычным сложением систем (8.21), (8.22), (8.23), в результате чего получается итоговая СЛАУ (8.16) с глобальной матрицей жесткости K и вектором f:

$$\begin{bmatrix} 28/9 & -53/18 & 0 & 0 \\ -53/18 & 56/9 & -53/18 & 0 \\ 0 & -53/18 & 56/9 & -53/18 \\ 0 & 0 & -53/18 & 28/9 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} -\left.\dfrac{d\widehat{u}}{dx}\right|_{x=0} \\ 0 \\ 0 \\ \left.\dfrac{d\widehat{u}}{dx}\right|_{x=1} \end{bmatrix}. \qquad (8.24)$$

В силу того что на границах $x = 0$ и $x = 1$ значения искомой функции известны из граничных условий первого рода (8.11), (8.12), первое и последнее уравнение в системе (8.24) можно исключить, положив в остальных уравнениях $u_1 = u(0) = 0$ и $u_4 = u(1) = 1$. Тогда из (8.24) имеем

$$\begin{bmatrix} 56/9 & -53/18 \\ -53/18 & 56/9 \end{bmatrix} \begin{bmatrix} u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 53/18 \end{bmatrix},$$

откуда $u_2 = 0{,}2885$; $u_3 = 0{,}6098$ (точные значения $u_2 = 0{,}2889$; $u_3 = 0{,}6102$).

Глобальная система (8.24) позволяет на границах, где заданы граничные условия первого рода, определить производные первого порядка. Из (8.24) получаем

$$-\frac{d\widehat{u}}{dx}\bigg|_{x=0} = \frac{28}{9}u_1 - \frac{53}{18}u_2,$$

$$\frac{d\widehat{u}}{dx}\bigg|_{x=1} = -\ \frac{53}{18}u_3 + \frac{28}{9}u_4,$$

откуда

$$(d\widehat{u}/dx)_{x=0} = (-53/18)\cdot 0{,}2885,$$

$$(d\widehat{u}/dx)_{x=1} = (-53/18)\cdot 0{,}6098 + (28/9)\cdot 1.$$

8.4.3. Случай граничных условий, содержащих производные. В случае, когда на границах расчетной области заданы граничные условия второго или третьего рода (или смешанные краевые условия), вместо слабой формулировки метода Галеркина (8.15) необходимо воспользоваться общей схемой взвешенных невязок (8.7) и на ее основе сформулировать слабую формулировку метода Галеркина.

Рассмотрим такую формулировку на примере следующей краевой задачи для ОДУ 2-го порядка:

$$\frac{d^2u}{dx^2} - u = 0, \quad 0 < x < 1; \tag{8.25}$$

$$u(0) = 0, \quad x = 0; \tag{8.26}$$

$$\frac{du(1)}{dx} = 1, \quad x = 1. \tag{8.27}$$

Тогда из (8.7) имеем

$$\int_0^1 W_s \left(\frac{d^2\widehat{u}}{dx^2} - \widehat{u}\right) dx + \left[\overline{W}_s \left(\frac{d\widehat{u}}{dx} - 1\right)\right]_{x=1} = 0, \tag{8.28}$$

где невязка на левой границе равна нулю, поскольку на ней задано значение искомой функции.

Для граничных условий, содержащих производные, базисные функции $\overline{W}_s$ на соответствующей границе можно подобрать так, чтобы эти производные и производные первого порядка, возникшие на той же границе от операции интегрирования по частям, сократились, что существенно упрощает решение.

Интегрируя первое слагаемое в (8.28) по частям, получим

$$-\int_0^1 \left(\frac{dW_s}{dx}\frac{d\widehat{u}}{dx} + W_s\widehat{u}\right) dx + W_s \left.\frac{d\widehat{u}}{dx}\right|_0^1 + \left[\overline{W}_s \left(\frac{d\widehat{u}}{dx} - 1\right)\right]_{x=1} = 0.$$

Положим здесь $\overline{W}_s\big|_{x=1} = -W_s|_{x=1}$ (а в соответствии с методом Галеркина $W_s = N_s$, $s = \overline{1, M+1}$), будем иметь

$$\int_0^1 \left(\frac{dN_s}{dx}\frac{d\widehat{u}}{dx} + N_s\widehat{u}\right) dx = -\left[N_s\frac{d\widehat{u}}{dx}\right]_{x=0} + [N_s]_{x=1}. \tag{8.29}$$

Проводя те же рассуждения, что и в задаче (8.10)–(8.12) с граничными условиями первого рода, из (8.29) получим следующую глобальную СЛАУ, в которой граничное условие на границе $x = 1$ учтено естественным образом ($[N_s]_{x=0} = 1$):

$$\begin{bmatrix} 28/9 & -53/18 & 0 & 0 \\ -53/18 & 56/9 & -53/18 & 0 \\ 0 & -53/18 & 56/9 & -53/18 \\ 0 & 0 & -53/18 & 28/9 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} -\left.\dfrac{d\widehat{u}}{dx}\right|_{x=0} \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{8.30}$$

Краевое условие на левой границе $u_1 = u(0) = 0$ учитывается вычеркиванием первого уравнения в (8.30) и равенством $u_1 = 0$ в остальных уравнениях.

Решая глобальную систему (8.24) или (8.30) относительно узловых значений искомой функции и подставляя их в линейную комбинацию (8.13), получим решение задачи методом конечных элементов не только в нумерованных узлах, но и в любой точке любого конечного элемента с помощью функции элемента.

§ 8.5. Метод конечных элементов в стационарных задачах математической физики

8.5.1. Основные этапы решения стационарных задач математической физики методом конечных элементов. Перечислим основные этапы решения двумерных стационарных задач математической физики с помощью МКЭ на основе метода Галеркина взвешенных невязок.

1. Расчетная область $\overline{\Omega} = \Omega + \Gamma \in R^2$, которая может быть и многосвязной, разбивается на элементы $\overline{\Omega}^e = \Omega^e + \Gamma^e$ того же пространства. Для двумерной расчетной области в качестве конечных элементов принимаются треугольные или четырехугольные элементы, причем последние можно разделить диагональю на два треугольных элемента. Достоинствами треугольных элементов являются возможность хорошей аппроксимации границы области и возможность аппроксимации искомой функции на треугольном элементе с помощью простейшей поверхности — плоскости, определяемой значениями искомой функции в нумерованных узлах элемента (обычно это вершины треугольного элемента). Разбиение на элементы должно удовлетворять условию $\bigcup\limits_e \overline{\Omega}^e = \overline{\Omega}$, причем смежные элементы должны иметь общие стороны и общие нумерованные узлы (рис. 8.8).

2. В нумерованных узлах фиксируются узловые значения искомой функции, являющиеся неизвестными величинами, подлежащими определению.

3. С помощью узловых значений в нумерованных узлах элемента искомая функция аппроксимируется поверхностью (чаще всего линейной функцией, описывающей плоскость), называемой *функцией элемента*, позволяющей определить искомую функцию в любой точке конечного элемента. При этом если

количество нумерованных узлов элемента на единицу больше размерности пространства R^n, то элемент называется *линейным*. Если в элементе число нумерованных узлов больше чем $n+1$, то этот элемент называется *нелинейным*, а искомая функция на нем аппроксимируется с помощью *нелинейной функции*.

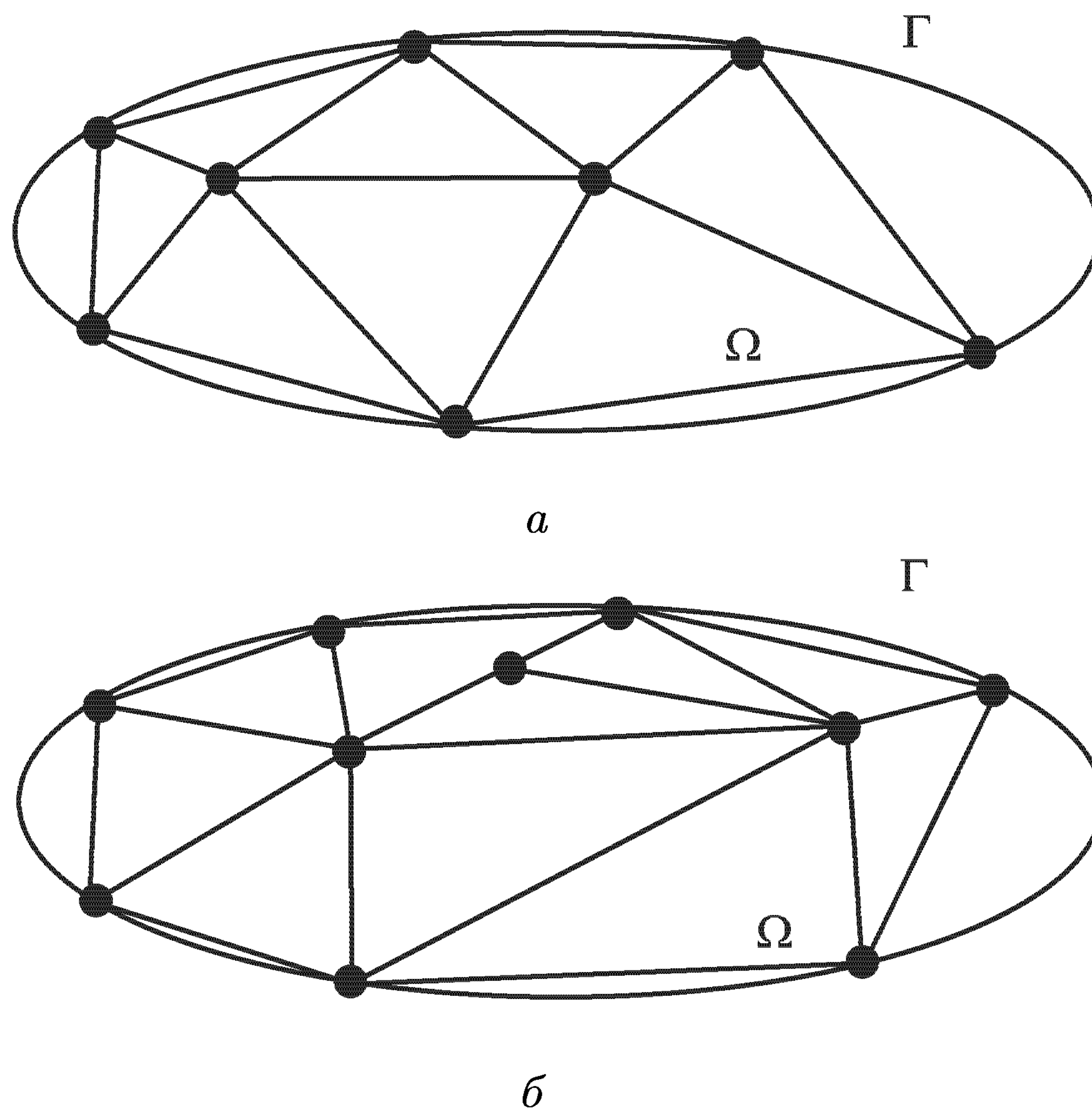


Рис. 8.8. Разбиение двумерной расчетной области: *а* — правильное разбиение; *б* — неправильное разбиение

4. Каждому нумерованному узлу расчетной области $\Omega + \Gamma$ приписывается базисная функция, равная единице в узле, которому она приписывается, и нулю во всех остальных узлах расчетной области. Базисные функции для различных нумерованных узлов являются линейно независимыми.

5. Приближенное решение задачи формируется в виде линейной комбинации базисных функций по всем нумерованным узлам расчетной области с коэффициентами линейной комбинации, равными узловым значениям искомой функции.

6. Это решение подставляется в дифференциальную задачу, что приводит не к тождественному нулю, поскольку подставляется приближенное решение, а к невязкам по расчетной области R_Ω и границе R_Γ.

7. В соответствии с различными методами взвешенных невязок, наиболее эффективным из которых является метод Галеркина, невязки ортогонализируются с системой весовых функций (в методе Галеркина в качестве весовых функций принимаются базисные функции). Результатом такой ортогонализации является глобальная система линейных алгебраических уравнений (СЛАУ) относительно узловых значений искомой функции. Число уравнений в этой системе совпадает с количеством нумерованных узлов расчетной области. Каждый элемент матрицы и вектора правых частей этой СЛАУ содержит в той или иной степени вклады элементов матриц и правых частей *локальных* СЛАУ, сформированных для каждого конечного элемента. Процесс суммирования таких вкладов конечных элементов называют *ансамблированием* конечных элементов, т. е. локальным номерам узлов конечных элементов ставятся в соответствие глобальные номера узлов расчетной области.

8. Решая полученную глобальную СЛАУ каким-либо из известных методов, получаем узловые значения искомой функции, с помощью которых из функций элементов определяются значения искомой функции в любых точках конечных элементов.

8.5.2. Принципы разбиения плоских областей на конечные элементы. Будем разбивать плоскую расчетную область $\Omega + \Gamma \in R^2$ на треугольные конечные элементы, при этом номер элемента обозначается верхним индексом e, а *локальные* номера узлов в элементе — нижними индексами i, j, k, причем если зафиксирован какой-либо узел под номером i, то остальные узлы под номерами j и k нумеруются против часовой стрелки.

Перечислим основные принципы разбиения плоских расчетных областей.

1. Сложные плоские области вначале разбивают на подобласти, которые затем разбивают на конечные элементы.

2. При разбиении на конечные элементы тонкостенных областей каждый конечный элемент должен содержать как минимум один нумерованный узел, находящийся внутри расчетной области, так как если все нумерованные узлы расположены на границе расчетной области, то при аппроксимации учитываются только граничные условия и не учитывается дифференциальное уравнение (рис. 8.9).

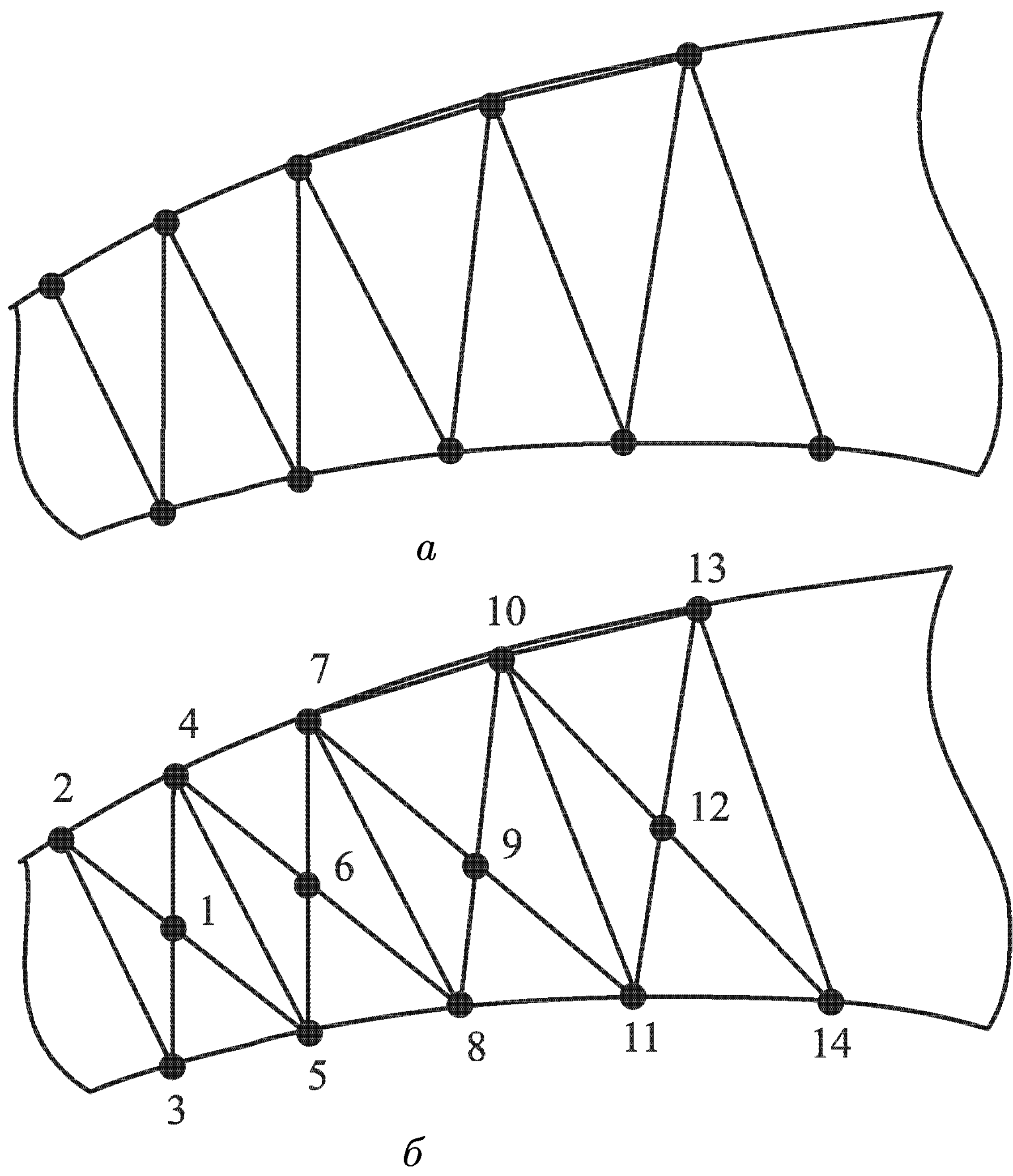


Рис. 8.9. Конечные элементы в тонкостенных расчетных областях: *а* — неправильное, *б* — правильное разбиение

3. Глобальную нумерацию узлов в расчетной области необходимо осуществлять таким образом, чтобы в отдельном конечном элементе разность между максимальным и минимальным номерами была как можно меньше, поскольку от этой разницы зависит полуширина b ленты матрицы (рис. 8.10) в глобальной

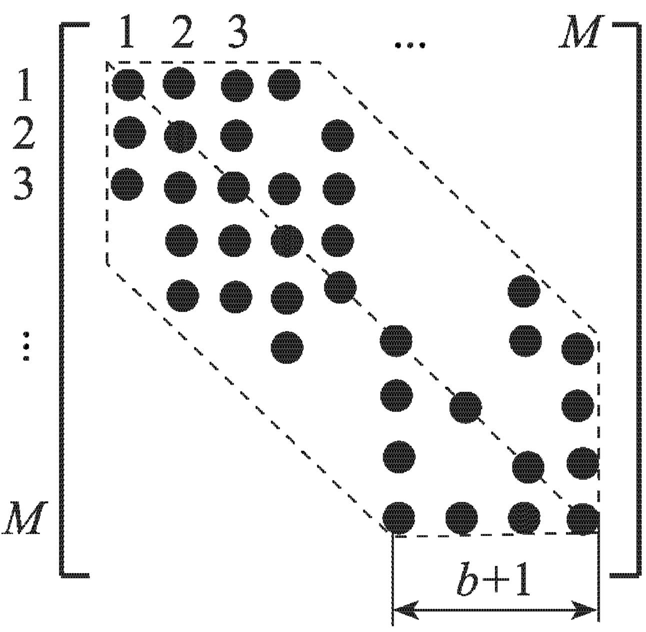


Рис. 8.10. Вид ленточной матрицы глобальной системы линейных алгебраических уравнений

матрице результирующей системы алгебраических уравнений, которая является *ленточной*, если все ее ненулевые элементы на полуширине b расположены около главной диагонали.

Чем меньше полуширина b ленточной матрицы, тем устойчивее решение СЛАУ. Таким образом, нумерация узлов в тонкостенной расчетной области, представленной на рис. 8.9, осуществляется чередованием номеров на верхней и нижней границах (нумерация узлов сначала по верхней, а затем по нижней границе приведет к значительной разности номеров в каждом конечном элементе и такому же увеличению полуширины ленты матрицы).

4. В многосвязных областях внутренние полости вписывают в многоугольники, в вершины которых помещают нумерованные узлы, соединяемые гранями с другими нумерованными узлами, расположенными внутри расчетной области.

8.5.3. Аппроксимация линейными многочленами и базисные функции. Применение метода конечных элементов рассмотрим на примере следующей третьей краевой задачи для уравнения Пуассона в области, представленной на рис. 8.11.

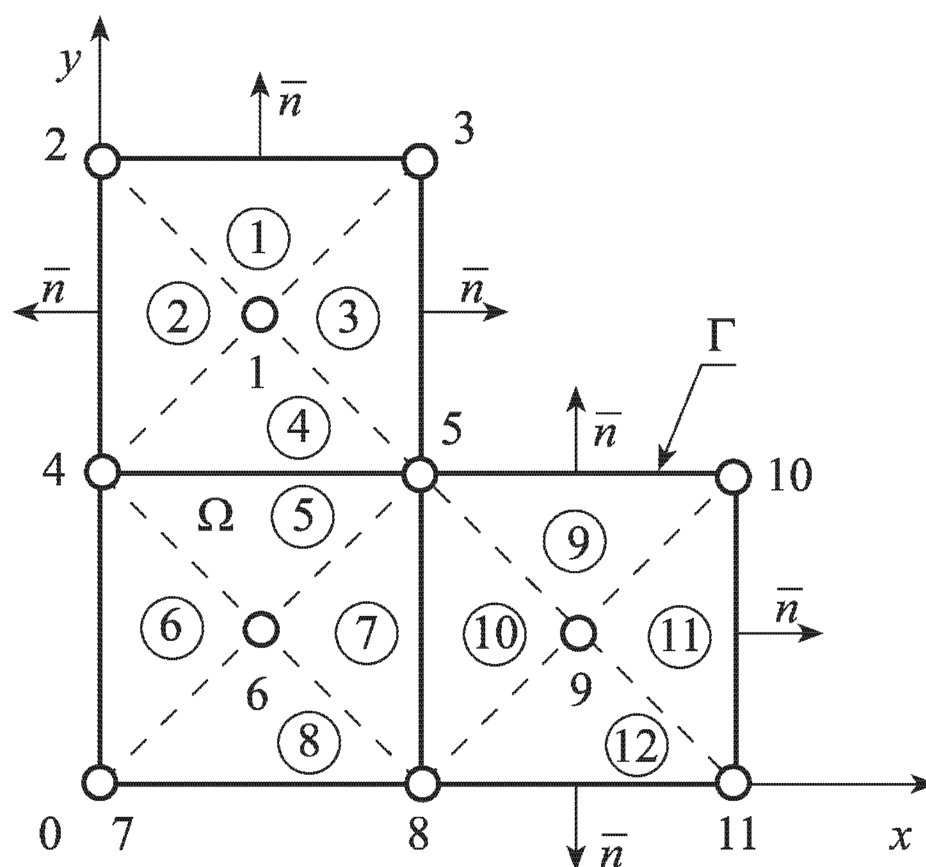


Рис. 8.11. Расчетная область

$$\frac{\partial}{\partial x}\left(\lambda(x,y)\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(\lambda(x,y)\frac{\partial u}{\partial y}\right)=f(x,y), \quad (x,y)\in\Omega\,; \tag{8.31}$$

$$\lambda(x, y) \left.\frac{\partial u}{\partial n}\right|_\Gamma + \alpha\, u|_\Gamma = \varphi(x, y), (x, y) \in \Gamma. \tag{8.32}$$

Разобьем вначале область Ω на три подобласти в виде трех четырехугольников. Затем внутри каждого четырехугольника введем по одному внутреннему узлу, которые соединим с узлами в вершинах четырехугольников. Пронумеровав все узлы, получим разбиение всей области на двенадцать треугольных конечных элементов с общим числом нумерованных узлов, равным одиннадцати.

Каждому нумерованному узлу приписывается базисная функция $N_m(x, y)$ в виде линейного многочлена, значение которого равно единице в узле, которому она приписана, и нулю — во

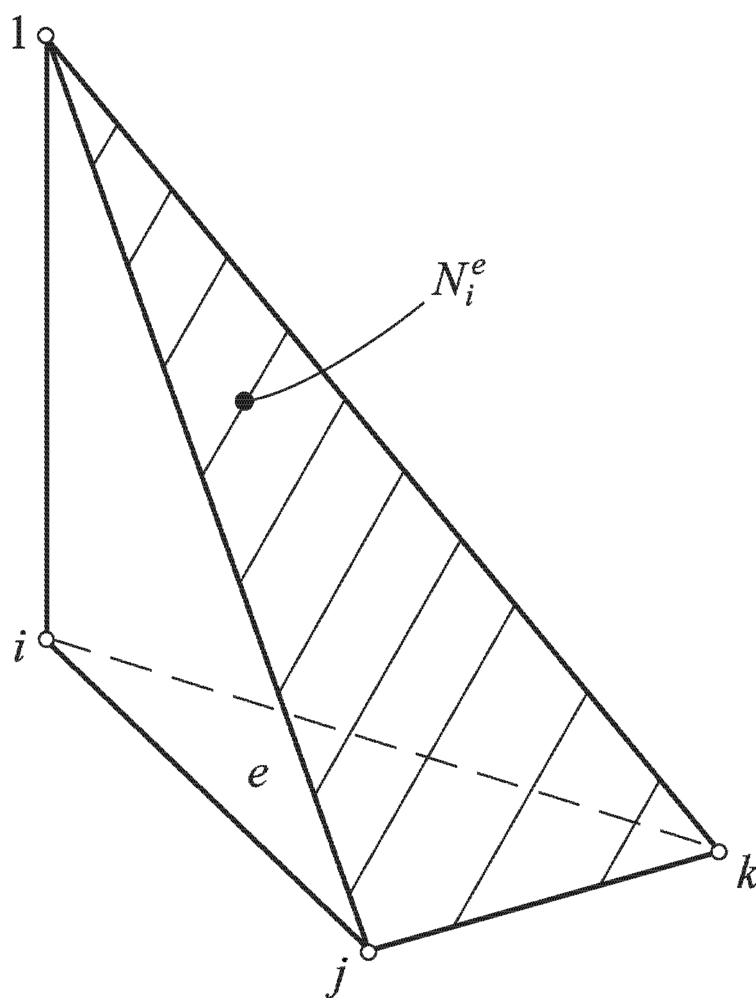


Рис. 8.12. Базисная функция N_i^e

всех остальных узлах расчетной области. На рис. 8.12 приведен пример базисной функции $N_i^e(x, y)$, приписанной узлу i в локальной нумерации узлов в конечном элементе e. Такая базисная функция имеет вид

$$N_i^e(x, y) = \alpha_i^e + \beta_i^e x + \gamma_i^e y, \tag{8.33}$$

где коэффициенты α_i^e, β_i^e, γ_i^e находятся из условий $N_i^e(x_i, y_i) = 1$, $N_i^e(x_j, y_j) = 0$, $N_i^e(x_k, y_k) = 0$. При этом координаты нумерованных узлов определены положением узлов в расчетной области относительно начала координат. Выполнение этих трех

условий приводит к следующей системе из трех линейных алгебраических уравнений относительно коэффициентов α_i^e, β_i^e, γ_i^e:

$$\begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{bmatrix} \begin{bmatrix} \alpha_i^e \\ \beta_i^e \\ \gamma_i^e \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \tag{8.34}$$

решением которой являются выражения

$$\alpha_i^e = \frac{1}{2\Delta^e}\left(x_j y_k - x_k y_j\right), \quad \beta_i^e = \frac{1}{2\Delta^e}\left(y_j - y_k\right),$$

$$\gamma_i^e = \frac{1}{2\Delta^e}\left(x_k - x_j\right),$$

где Δ^e — площадь элемента,

$$\Delta^e = \frac{1}{2}\det\begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{bmatrix}.$$

Для остальных узлов j и k конечного элемента e базисные функции $N_j^e(x,y)$ и $N_k^e(x,y)$ формируются аналогично. Тогда равенство (8.33) для этих узлов будет иметь вид

$$N_j^e(x,y) = \alpha_j^e + \beta_j^e x + \gamma_j^e y,$$

$$N_k^e(x,y) = \alpha_k^e + \beta_k^e x + \gamma_k^e y,$$

а система (8.34) сохраняется, за исключением двух групп неизвестных α_j^e, β_j^e, γ_j^e; α_k^e, β_k^e, γ_k^e, и двух векторов правых частей $[\,0 \;\; 1 \;\; 0\,]^T$ для N_j^e и $[\,0 \;\; 0 \;\; 1\,]^T$ для N_k^e.

Приближенное решение задачи (8.31), (8.32) находится в виде линейной комбинации базисных функций, коэффициентами которой являются значения искомой функции в нумерованных узлах, т. е. в форме

$$u(x,y) \approx \widehat{u}(x,y) = \sum_{m=1}^{M} u_m N_m(x,y), \tag{8.35}$$

где M — число нумерованных узлов.

Для отдельного элемента e решение (8.35) представляется в виде

$$u^e\,(x,y) = u_i^e(x,y)N_i^e(x,y) + \\ + u_j^e(x,y)N_j^e(x,y) + u_k^e(x,y)N_k^e(x,y) \quad (8.36)$$

и называется *функцией элемента*. С ее помощью значение искомой функции можно определить в любой точке конечного элемента, как только станут известными узловые значения u_i^e, u_j^e, u_k^e. Система базисных функций $N_m(x,y)$, $m = \overline{1, M}$, обладает свойством полноты, поскольку при $M \to \infty$ решение (8.35) может сколь угодно точно аппроксимировать искомую функцию.

8.5.4. Слабая формулировка конечно-элементного метода Галеркина. Если подставить приближенное решение (8.35) в дифференциальную задачу (8.31), (8.32), то результатом подстановки будет не тождественный нуль, поскольку (8.35) — приближенное решение, а некоторая функциональная невязка $R_\Omega(x,y)$ по расчетной области Ω и невязка $R_\Gamma(x,y)$ — по границе Γ:

$$R_\Omega(x,y) = \frac{\partial}{\partial x}\left(\lambda\frac{\partial\widehat{u}}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda\frac{\partial\widehat{u}}{\partial y}\right) - f(x,y),$$

$$R_\Gamma(x,y) = \lambda\left.\frac{\partial\widehat{u}}{\partial n}\right|_\Gamma + \alpha\widehat{u}\big|_\Gamma - \varphi(x,y).$$

В соответствии с методами взвешенных невязок требуем ортогональности этих функциональных невязок и специальным образом подобранных весовых функций $W_s(x,y)$, $s = \overline{1, M}$, для невязки $R_\Omega(x,y)$ и $\overline{W}_s(x,y)$, $s = \overline{1, M}$, для невязки $R_\Gamma(x,y)$. Для непрерывных функций $R_\Omega(x,y)$, $R_\Gamma(x,y)$ это означает равенство нулю скалярных произведений $(R_\Omega, W_s) = 0$ и $\left(R_\Gamma, \overline{W}_s\right) = 0$ или $(R_\Omega, W_s) + \left(R_\Gamma, \overline{W}_s\right) = 0$, $s = \overline{1, M}$, что приводит к равенству нулю суммы двойного и криволинейного интегралов соответственно по области Ω и границе Γ:

$$(R_\Omega, W_s) + \left(R_\Gamma, \overline{W}_s\right) = \\ = \iint\limits_\Omega \left[\frac{\partial}{\partial x}\left(\lambda\frac{\partial\widehat{u}}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda\frac{\partial\widehat{u}}{\partial y}\right) - f(x,y)\right] W_s(x,y)dxdy +$$

$$+\int\limits_{\Gamma}\left[\lambda\left.\frac{\partial\widehat{u}}{\partial n}\right|_{\Gamma}+a\widehat{u}\big|_{\Gamma}-\varphi(x,y)\right]\overline{W}_s d\Gamma=0,\; s=\overline{1,M}, \quad (8.37)$$

где $W_s(x,y)$ — весовые функции для внутренних узлов расчетной области, а $\overline{W}_s(x,y)$ — весовые функции для граничных узлов расчетной области, $s=\overline{1,M}$. Весовые функции выбираются таким образом, чтобы они были ортогональны невязкам по области Ω и по границе Γ. В соответствии с методом Галеркина взвешенных невязок весовые функции равны базисным: $W_s = N_s$, $\overline{W}_s = \overline{N}_s$.

Для базисных функций вида (8.33) интегралы от вторых производных базисных функций (в соответствии с (8.35) приближенное решение $\widehat{u}(x,y)$ содержит базисные функции) не существуют, поскольку они стремятся к $+\infty$ или $-\infty$. Для ослабления гладкости подынтегральных функций в (8.37) используем первую формулу Грина, согласно которой для двух достаточное число раз непрерывно дифференцируемых функций $u(x,y)$ и $v(x,y)$ имеет место равенство

$$\int\limits_{\Omega} v\left[\frac{\partial}{\partial x}\left(\lambda\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(\lambda\frac{\partial u}{\partial y}\right)\right]dxdy=$$

$$=-\int\limits_{\Omega}\left[\frac{\partial v}{\partial x}\cdot\lambda\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\cdot\lambda\frac{\partial u}{\partial y}\right]dxdy+\int\limits_{\Gamma} v\cdot\lambda\frac{\partial u}{\partial n}d\Gamma,$$

являющееся обобщением на двумерный случай формулы интегрирования по частям в определенном интеграле.

Тогда, подставляя в (8.37) решение (8.35) и применяя к полученному выражению первую формулу Грина, в которой вместо функции $v(x,y)$ подставлена функция $N_s(x,y)$, а вместо $u(x,y)$ — функция $\widehat{u}(x,y)$, получим слабую формулировку конечно-элементного метода Галеркина:

$$\sum_{m=1}^{M} u_m\cdot\left\{-\int\limits_{\Omega}\left[\frac{\partial N_s}{\partial x}\cdot\lambda\frac{\partial N_m}{\partial x}+\frac{\partial N_s}{\partial y}\cdot\lambda\frac{\partial N_m}{\partial y}\right]dxdy+\right.$$

$$\left.+\int\limits_{\Gamma} N_s\cdot\lambda\frac{\partial N_m}{\partial n}d\Gamma+\int\limits_{\Gamma}\overline{N}_s\cdot\lambda\frac{\partial N_m}{\partial n}d\Gamma+\int\limits_{\Gamma}\overline{N}_s\cdot\alpha N_m d\Gamma\right\}=$$

$$= \int\limits_{\Omega} f(x,y) N_s dx dy + \int\limits_{\Gamma} \overline{N}_s \varphi(x,y) d\Gamma, \quad s = \overline{1,M}. \quad (8.38)$$

Весовые функции $\overline{N}_s$ на границе Γ можно выбрать таким образом, чтобы криволинейные интегралы по границе (второй и третий интегралы в левой части выражения (8.38)) сократились, т. е. принять $\overline{N}_s = -N_s$. Тогда выражение (8.38) примет вид

$$\sum_{m=1}^{M} u_m \times$$

$$\times \left\{ \int\limits_{\Omega} \left[\frac{\partial N_s}{\partial x} \cdot \lambda \frac{\partial N_m}{\partial x} + \frac{\partial N_s}{\partial y} \cdot \lambda \frac{\partial N_m}{\partial y} \right] dx dy + \alpha \int\limits_{\Gamma} N_s \cdot N_m d\Gamma \right\} =$$

$$= -\int\limits_{\Omega} f(x,y) N_s dx dy + \int\limits_{\Gamma} N_s \varphi(x,y) d\Gamma, \quad s = \overline{1,M}. \quad (8.39)$$

Выражения (8.39) — это неоднородная система линейных алгебраических уравнений порядка M, которая в векторно-матричной форме имеет вид

$$Au = F, \quad (8.40)$$

где элементы a_{sm} матрицы A и правых частей F образованы суммированием вкладов отдельных конечных элементов (в соответствии с аддитивным свойством кратных и криволинейных интегралов):

$$a_{sm} = \sum_{e=1}^{E} a_{sm}^e, \quad F_s = \sum_{e=1}^{E} F_s^e.$$

СЛАУ (8.39) или (8.40) называют *глобальной* СЛАУ, матрицу A и вектор правых частей F — соответственно *глобальной матрицей* (или матрицей жесткости) и *глобальным вектором правых частей.*

Для отдельного конечного элемента $\Omega^e + \Gamma^e$ с локальными номерами узлов i, j, k на основе глобальной СЛАУ (8.40), элементы матрицы и вектора правых частей которой определяются выражением (8.39), можно построить *локальную* СЛАУ

$$A^e u^e = F^e \quad (8.41)$$

с *локальной матрицей* $A^e = [a^e_{sm}]$, $s = i, j, k$; $m = i, j, k$ и *локальным вектором правых частей* $F^e = [f^e_s]$, $s = i, j, k$, т. е.

$$A^e = \begin{bmatrix} a^e_{ii} & a^e_{ij} & a^e_{ik} \\ a^e_{ji} & a^e_{jj} & a^e_{jk} \\ a^e_{ki} & a^e_{kj} & a^e_{kk} \end{bmatrix}, \quad F^e = \begin{bmatrix} F^e_i \\ F^e_j \\ F^e_k \end{bmatrix}.$$

Элементы a^e_{sm}, $s = i, j, k; m = i, j, k$, имеют вид выражений в фигурных скобках СЛАУ (8.39), где интегралы вычисляются по конечному элементу $\Omega^e + \Gamma^e$, а компоненты F^e_s, $s = i, j, k$ имеют вид правых частей СЛАУ (8.39) на конечном элементе $\Omega^e + \Gamma^e$, т. е.

$$a^e_{sm} = \iint\limits_{\Omega^e} \left(\frac{\partial N^e_s}{\partial x} \lambda \frac{\partial N^e_m}{\partial x} + \frac{\partial N^e_s}{\partial y} \lambda \frac{\partial N^e_m}{\partial y} \right) dxdy + \\ + \alpha \int\limits_{\Gamma^e} N^e_s N^e_m d\Gamma, \quad s = i, j, k; \quad m = i, j, k; \quad (8.42)$$

$$F^e_s = \int\limits_{\Gamma^e} N^e_s \varphi^e(x, y) d\Gamma - \iint\limits_{\Omega^e} f^e(x, y) N^e_s dxdy, \quad s = i, j, k. \quad (8.43)$$

В выражениях (8.42), (8.43) криволинейные интегралы равны нулю, если конечный элемент не содержит граней, являющихся граничными. Элемент a^e_{sm} локальной матрицы A^e вычисляется таким образом, что для каждого номера s перебираются все номера m. Например, для локального номера $s = i$ перебираются все номера $m = i, j, k$; для $s = j$, $m = i, j, k$; для $s = k$, $m = = i, j, k$.

При вычислении двойных интегралов в выражениях (8.42), (8.43) используются базисные функции (8.33) и теорема о среднем

$$\iint\limits_{\Omega^e} \left(\frac{\partial N^e_s}{\partial x} \lambda^e(x, y) \frac{\partial N^e_m}{\partial x} + \frac{\partial N^e_s}{\partial y} \lambda^e(x, y) \frac{\partial N^e_m}{\partial y} \right) dxdy = \\ = \lambda^e_{\text{ср}} \left(\beta^e_s \beta^e_m + \gamma^e_s \gamma^e_m \right) \cdot \Delta^e, \quad s, m = i, j, k; \quad (8.44)$$

$$\iint\limits_{\Omega^e} f^e(x,y)N_s^e dxdy = f_{\text{ср}}^e \cdot \frac{1}{3}\Delta^e, \quad s = i,j,k, \tag{8.45}$$

где в соответствии с теоремой о среднем

$$\lambda_{\text{ср}}^e = \lambda\left(x^*, y^*\right), \quad f_{\text{ср}}^e = f\left(x^*, y^*\right),$$

$$x^* = \frac{1}{3}\left(x_i + x_j + x_k\right), \quad y^* = \frac{1}{3}\left(y_i + y_j + y_k\right),$$

Δ^e — площадь треугольного элемента, β_s^e, γ_s^e — коэффициенты в базисных функциях (8.33).

При вычислении криволинейных интегралов в выражениях (8.42), (8.43) положим для определенности, что на границу Г выходят узлы i и k элемента $\Omega^e + \Gamma^e$ (рис. 8.13).

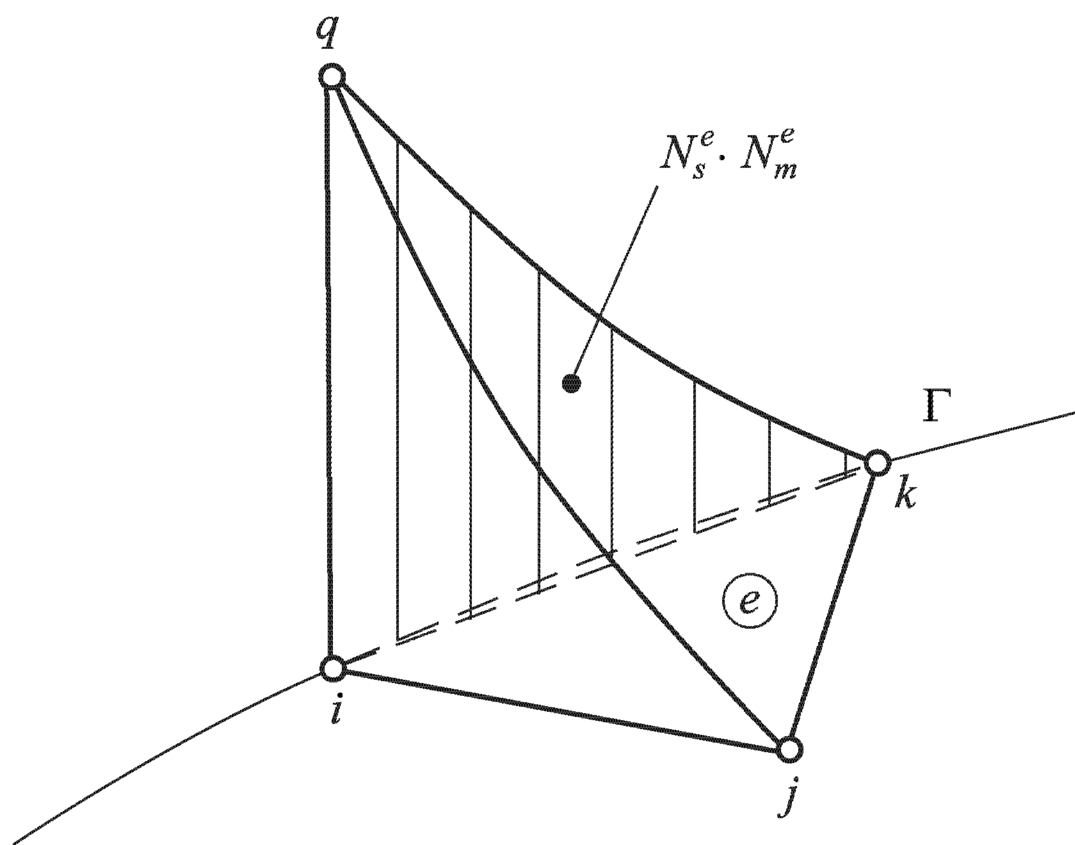


Рис. 8.13. К вычислению криволинейных интегралов

Тогда произведения $N_s^e(x,y) \cdot N_m^e(x,y)$, s, $m = i,k$, геометрически представляют собой поверхности второго порядка, а их сечения плоскостью qik являются кривыми второго порядка, а точнее — квадратными параболами. Таким образом, криволинейный интеграл в выражении (8.42) геометрически равен заштрихованной площади на рис. 8.13, т. е.

$$\int\limits_{\Gamma^e} N_s^e N_m^e d\Gamma = \frac{1}{3}\sqrt{\left(x_k - x_i\right)^2 + \left(y_k - y_i\right)^2}, \quad s = m, \tag{8.46}$$

$$\int\limits_{\Gamma^e} N_s^e N_m^e d\Gamma = \frac{1}{6}\sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}, \quad s \neq m, \tag{8.47}$$

поскольку площадь криволинейного треугольника qik в соответствии с теоремой о среднем равна $1/3\Delta\Gamma$, где $\Delta\Gamma$ — расстояние между узлами i и k.

По той же причине имеет место равенство

$$\int\limits_{\Gamma^e} N_s^e \cdot \varphi^e(x, y)\, d\Gamma = \varphi(x^*, y^*) \cdot \frac{1}{2}\sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}, \tag{8.48}$$

где $x^* = \frac{1}{2}(x_i + x_k)$; $y^* = \frac{1}{2}(y_i + y_k)$, $\varphi_{\text{ср}}^e = \varphi(x^*, y^*)$.

Таким образом, в соответствии с выражениями (8.44)–(8.48) локальная матрица A^e и вектор правых частей F^e СЛАУ для конечного элемента $\overline{\Omega}^e = \Omega^e + \Gamma^e$ имеют вид (в случае, когда на границу Γ выходят узлы i и k конечного элемента)

$$A^e = \lambda_{\text{ср}}^e \cdot \Delta^e \cdot \begin{bmatrix} (\beta_i^e)^2 + (\gamma_i^e)^2 & \beta_i^e \beta_j^e + \gamma_i^e \gamma_j^e & \beta_i^e \beta_k^e + \gamma_i^e \gamma_k^e \\ \beta_j^e \beta_i^e + \gamma_j^e \gamma_i^e & \left(\beta_j^e\right)^2 + \left(\gamma_j^e\right)^2 & \beta_j^e \beta_k^e + \gamma_j^e \gamma_k^e \\ \beta_k^e \beta_i^e + \gamma_k^e \gamma_i^e & \beta_k^e \beta_j^e + \gamma_k^e \gamma_j^e & (\beta_k^e)^2 + (\gamma_k^e)^2 \end{bmatrix} +$$

$$+ \frac{\alpha}{3}\sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 0 & 0 \\ 1/2 & 0 & 1 \end{bmatrix}, \tag{8.49}$$

$$F^e = \begin{bmatrix} \varphi_{\text{ср}}^e \cdot \frac{1}{2}\sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} - f_{\text{ср}}^e \cdot \Delta^e/3 \\ -f_{\text{ср}}^e \cdot \Delta^e/3 \\ \varphi_{\text{ср}}^e \cdot \frac{1}{2}\sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} - f_{\text{ср}}^e \cdot \Delta^e/3 \end{bmatrix}. \tag{8.50}$$

В случае, если решается первая краевая задача для того же уравнения (8.31), слабая формулировка (8.38) примет вид $\left(\overline{W}_s = N_s\right)$

$$\sum_{m=1}^{M} u_m \left\{ \iint_{\Omega} \left[\frac{\partial N_s}{\partial x} \lambda \frac{\partial N_m}{\partial x} + \frac{\partial N_s}{\partial y} \lambda \frac{\partial N_m}{\partial y} \right] dxdy - \right.$$

$$\left. - \int_{\Gamma} N_s \cdot \lambda \frac{\partial N_m}{\partial n} d\Gamma - \int_{\Gamma} N_s N_m d\Gamma \right\} =$$

$$= - \int_{\Omega} N_s f\,(x,y) dxdy - \int_{\Gamma} N_s \varphi\,(x,y) d\Gamma, \quad s = \overline{1, M}, \quad (8.51)$$

где

$$\frac{\partial N_m}{\partial n} = \frac{\partial N_m}{\partial x} \cos(\mathbf{i}, \mathbf{n}) + \frac{\partial N_m}{\partial y} \cos(\mathbf{j}, \mathbf{n}),$$

n — вектор внешней нормали к границе Г.

Поэтому к вычислению интегралов (8.44)–(8.48) необходимо добавить вычисление криволинейных интегралов от произведения весовых функций на производные по нормали от базисных функций. Тогда

$$\int_{\Gamma^e} N_s^e \cdot \lambda \frac{\partial N_m^e}{\partial n} d\Gamma =$$

$$= \lambda_{\text{ср}}^e \int_{\Gamma^e} \left[N_s^e \frac{\partial N_m^e}{\partial x} \cos(\mathbf{i}, \mathbf{n}) + N_s^e \frac{\partial N_m^e}{\partial y} \cos(\mathbf{j}, \mathbf{n}) \right] d\Gamma =$$

$$= \lambda_{\text{ср}}^e \cdot [\beta_m^e \cos(\mathbf{i}, \mathbf{n}) + \gamma_m^e \cos(\mathbf{j}, \mathbf{n})] \cdot \frac{1}{2} \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}, \quad (8.52)$$

если на границу Г выходят узлы i и k.

Если конечные элементы ни одной гранью не выходят на границу области Г, то соответствующие криволинейные интегралы равны нулю.

8.5.5. Ансамблирование элементов и построение глобальной СЛАУ. Для построения глобальной матрицы жесткости A и вектора правых частей F системы (8.40) необходимо локальным номерам узлов i, j, k каждого элемента поставить в соответствие глобальные номера узлов m, $m = \overline{1, M}$. При этом

каждая локальная матрица A^e элемента расширяется до размера $M \times M$, т. е. в глобальной матрице A ненулевые элементы локальной матрицы A^e становятся на места, определяемые глобальными номерами узлов элемента Ω^e. Аналогично для вектора правых частей F.

Затем расширенные таким образом матрицы и векторы правых частей всех конечных элементов складываются, в результате чего получаем глобальную матрицу A и вектор правых частей F, или глобальную СЛАУ (8.40).

Такой процесс объединения локальных СЛАУ для конечных элементов в глобальную СЛАУ называется *ансамблированием конечных элементов*.

Для рассматриваемого примера (рис. 8.11) $M = 11$. Тогда, например, для элемента с номером 4 определено следующее соответствие локальных и глобальных номеров узлов: $i = 1$, $j = 4$, $k = 5$. Тогда локальная матрица элемента 4, имеющая вид

$$A^4 = \begin{bmatrix} a_{ii}^4 & a_{ij}^4 & a_{ik}^4 \\ a_{ji}^4 & a_{jj}^4 & a_{jk}^4 \\ a_{ki}^4 & a_{kj}^4 & a_{kk}^4 \end{bmatrix},$$

преобразуется в расширенную матрицу $A^4_{\text{рас}}$ размера 11×11 следующим образом:

$$A^4_{\text{рас}} = \begin{array}{c|c} & m \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \ \ldots \ 11 \\ \hline s & \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \\ 11 \end{array} & \begin{bmatrix} a_{11}^4 & 0 & 0 & a_{14}^4 & a_{15}^4 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ a_{41}^4 & 0 & 0 & a_{44}^4 & a_{45}^4 & 0 & \ldots & 0 \\ a_{51}^4 & 0 & 0 & a_{54}^4 & a_{55}^4 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ \multicolumn{8}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \end{bmatrix} \end{array}$$

Аналогично вектор правых частей $F^4 = \left[F_i^4, F_j^4, F_k^4\right]^T$ локальной системы (8.41) преобразуется в расширенный вектор с одиннадцатью компонентами:

$$F_{\text{рас}}^4 = \left[F_1^4 \; 0 \; 0 \; F_4^4 \; F_5^4 \;\; 0 \, \ldots \, 0\right]^T. \tag{8.53}$$

Операция суммирования осуществляется в соответствии с аддитивным свойством кратных и криволинейных интегралов. При этом в памяти достаточно хранить матрицу-сумматор и вектор-сумматор, а также буферные матрицу и вектор текущего элемента. Результирующую СЛАУ (8.40) можно решить одним из известных методов, а значения искомой функции во внутренних точках конечного элемента определяются по узловым значениям с помощью функции элемента (8.36).

§ 8.6. Метод конечных элементов в многомерных нестационарных задачах математической физики

В нестационарных задачах математической физики искомая функция зависит не только от пространственных переменных, но и от времени, а дифференциальные уравнения и начальные условия могут содержать производные по времени. Конечно-элементную аппроксимацию в таких задачах можно осуществить по пространственным переменным, сохранив дифференциальные операторы по времени, в результате чего получается система обыкновенных дифференциальных уравнений порядка M (M — число нумерованных узлов в пространственной расчетной области Ω), т. е. один из вариантов метода прямых (см. раздел 6.6.3).

Однако существует более простой и эффективный метод, согласно которому дифференциальный оператор по времени аппроксимируется конечно-разностным оператором, а пространственные операторы аппроксимируются конечно-элементным методом Галеркина. Рассмотрим этот метод на примере следующей двумерной нестационарной задачи теплопроводности с граничным условием 3-го рода для функции $u(x, y, t)$:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\lambda(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda(x,y)\frac{\partial u}{\partial y}\right) + f(x,y), \quad (x,y)\in\Omega, \quad t>0; \tag{8.54}$$

$$\lambda(x,y)\left.\frac{\partial u}{\partial n}\right|_{\Gamma} + \alpha u|_{\Gamma} = \varphi(x,y), (x,y)\in\Gamma, \quad t>0; \tag{8.55}$$

$$u(x,y,0) = \psi(x,y), (x,y)\in\overline{\Omega}, \quad t=0. \tag{8.56}$$

Решение задачи (8.55)–(8.57) представляется в виде линейной комбинации базисных функций $N_m(x,y)$, $m=\overline{1,M}$, с коэффициентами линейной комбинации, равными узловым значениям искомой функции $u_m(t)$, зависящими от времени:

$$u(x,y,t) \approx \widehat{u}(x,y,t) = \sum_{m=1}^{M} u_m(t)\cdot N_m(x,y), \tag{8.57}$$

где базисные функции определяются соотношениями вида (8.33).

Тогда в соответствии со слабой формулировкой конечно-элементного метода Галеркина (8.38) для задачи (8.54)–(8.56) можно записать следующее нестационарное выражение:

$$\sum_{m=1}^{M} u_m(t)\left\{\iint\limits_{\Omega}\left(\frac{\partial N_s}{\partial x}\lambda\frac{\partial N_m}{\partial x} + \frac{\partial N_s}{\partial y}\lambda\frac{\partial N_m}{\partial y}\right)dxdy + \right.$$
$$\left. + \int\limits_{\Gamma}\alpha N_s N_m d\Gamma\right\} + \sum_{m=1}^{M}\iint\limits_{\Omega} N_s\cdot N_m\frac{du_m(t)}{dt}dxdy =$$
$$= \iint\limits_{\Omega} N_s\cdot f(x,y)dxdy + \int\limits_{\Gamma} N_s\varphi(x,y)d\Gamma, s=\overline{1,M}, \tag{8.58}$$

в котором функцию $u_m(t)$ будем аппроксимировать на верхнем временном слое, обозначив $u_m(t^{k+1}) \equiv u_m^{k+1}$, а производную $\frac{du_m(t)}{dt}$ — с помощью отношения конечных разностей по времени справа:

$$\frac{du_m(t)}{dt} = \frac{u_m^{k+1} - u_m^k}{\tau} + O(\tau).$$

Тогда (8.58) будет иметь вид

$$\sum_{m=1}^{M} u_m^{k+1} \left\{ \iint_{\Omega} \left(\frac{\partial N_s}{\partial x} \lambda \frac{\partial N_m}{\partial x} + \frac{\partial N_s}{\partial y} \lambda \frac{\partial N_m}{\partial y} + \frac{1}{\tau} N_s N_m \right) dxdy + \right.$$

$$\left. + \int_{\Gamma} \alpha N_s N_m d\Gamma \right\} = \sum_{m=1}^{M} \iint_{\Omega} N_s \cdot N_m \frac{u_m^k}{\tau} dxdy +$$

$$+ \iint_{\Omega} N_s \cdot f(x,y) dxdy + \int_{\Gamma} N_s \varphi(x,y) d\Gamma, s = \overline{1, M}. \quad (8.59)$$

Систему линейных алгебраических уравнений (8.59) относительно неизвестных узловых значений u_m^{k+1}, $m = \overline{1, M}$, на верхнем временном слое в методе конечных элементов получаем с помощью аддитивного свойства кратных и криволинейных интегралов путем суммирования локальных СЛАУ (8.41) для отдельных конечных элементов с локальными номерами узлов i, j и k. При этом элементы локальной матрицы и вектора правых частей для конечного элемента $\overline{\Omega}^e$ выписываются на основе (8.59) следующим образом:

$$a_{sm}^e = \iint_{\Omega^e} \left(\frac{\partial N_s^e}{\partial x} \lambda^e \frac{\partial N_m^e}{\partial x} + \frac{\partial N_s^e}{\partial y} \lambda^e \frac{\partial N_m^e}{\partial y} + \frac{1}{\tau} N_s^e N_m^e \right) dxdy +$$

$$+ \int_{\Gamma^e} \alpha N_s^e N_m^e d\Gamma, \quad s, m = i, j, k. \quad (8.60)$$

$$F_s^e = \iint_{\Omega^e} N_s^e \cdot N_m^e \frac{u_m^k}{\tau} dxdy + \iint_{\Omega^e} N_s^e \cdot f^e(x,y) dxdy +$$

$$+ \int_{\Gamma^e} N_s^e \varphi^e(x,y) d\Gamma, \quad s = \overline{1, M}. \quad (8.61)$$

Кратные и криволинейные интегралы в выражениях (8.60), (8.61) вычисляются путем использования выражений согласно (8.44)–(8.52). В (8.61) u_m^k, как константа, выносится за знак интеграла.

В остальном алгоритм тот же, что и для стационарных задач. Исключение составляет тот факт, что глобальную СЛАУ приходится решать на каждом временном слое.

§ 8.7. Особенности решения пространственных задач математической физики методом конечных элементов

Рассмотрим эти особенности на примере следующей третьей краевой задачи для трехмерного уравнения Пуассона в области $\overline{\Omega} \in R^3$:

$$\frac{\partial}{\partial x}\left(\lambda\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial z}\left(\lambda\frac{\partial u}{\partial z}\right) = \\ = f(x, y, z), \ (x, y, z) \in \Omega; \quad (8.62)$$

$$\lambda \left.\frac{\partial u}{\partial n}\right|_{\Gamma} + au|_{\Gamma} = \varphi(x, y, z), \quad (x, y, z) \in \Gamma. \quad (8.63)$$

Приближенное решение ищется в виде следующей линейной комбинации базисных функций $N_m(x, y, z)$, $m = \overline{1, M}$:

$$u(x, y, z) \approx \widehat{u}(x, y, z) = \sum_{m=1}^{M} u_m \cdot N_m(x, y, z), \quad (8.64)$$

где коэффициентами линейной комбинации являются значения u_m, $m = \overline{1, M}$, искомой функции в нумерованных узлах трехмерных конечных элементов Ω^e.

В трехмерном случае в качестве конечных элементов принимаются *тетраэдры* с нумерованными узлами i, j, k, l в вершинах или *параллелепипеды* с узлами i, j, k, l, m, n, o, p. Тетраэдр является линейным конечным элементом, поскольку число нумерованных узлов на единицу больше размерности пространства R^3.

Если расчетная область $\overline{\Omega}$ разбита на тетраэдры, то базисные функции, ассоциируемые с каждым нумерованным узлом конечного элемента Ω^e, формируются в виде линейных функций переменных x, y, z, удовлетворяющих условиям равенства единице в узлах, для которых они определены, и нулю — в остальных

узлах. Для узла i, например, базисная функция имеет вид

$$N_i^e(x,y,z) = \alpha_i^e + \beta_i^e x + \gamma_i^e y + \delta_i^e z, \tag{8.65}$$

причем

$$N_i^e(x,y,z) = \begin{cases} 1 \text{ в узле } i; \\ 0 \text{ в узле } j,k,l. \end{cases}$$

Аналогично определяются базисные функции в узлах j, k, l.

С учетом этих требований к базисным функциям формируется следующая СЛАУ для определения коэффициентов α_i^e, β_i^e, γ_i^e, δ_i^e базисной функции $N_i^e(x,y,z)$ (8.65):

$$\begin{bmatrix} 1 & x_i & y_i & z_i \\ 1 & x_j & y_j & z_j \\ 1 & x_k & y_k & z_k \\ 1 & x_l & y_l & z_l \end{bmatrix} \begin{bmatrix} \alpha_i^e \\ \beta_i^e \\ \gamma_i^e \\ \delta_i^e \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{8.66}$$

откуда

$$\alpha_i^e = \frac{\Delta_\alpha^e}{6\Delta^e}, \quad \beta_i^e = \frac{\Delta_\beta^e}{6\Delta^e}, \quad \gamma_i^e = \frac{\Delta_\gamma^e}{6\Delta^e}, \quad \delta_i^e = \frac{\Delta_\delta^e}{6\Delta^e};$$

$$\Delta_\alpha^e = \det \begin{bmatrix} x_j & y_j & z_j \\ x_k & y_k & z_k \\ x_l & y_l & z_l \end{bmatrix}, \quad \Delta_\beta^e = -\det \begin{bmatrix} 1 & y_j & z_j \\ 1 & y_k & z_k \\ 1 & y_l & z_l \end{bmatrix},$$

$$\Delta_\gamma^e = \det \begin{bmatrix} 1 & x_j & z_j \\ 1 & x_k & z_k \\ 1 & x_l & z_l \end{bmatrix}, \quad \Delta_\delta^e = -\det \begin{bmatrix} 1 & x_j & y_j \\ 1 & x_k & y_k \\ 1 & x_l & y_l \end{bmatrix},$$

$$6\Delta^e = \det \begin{bmatrix} 1 & x_i & y_i & z_i \\ 1 & x_j & y_j & z_j \\ 1 & x_k & y_k & z_k \\ 1 & x_l & y_l & z_l \end{bmatrix}.$$

Здесь Δ^e — объем элемента Ω^e (как известно, объем тетраэдра равен шестой части объема параллелепипеда, который, в свою очередь, равен модулю смешанного произведения трех векторов, исходящих из одной точки, например из узла i в узлы j, k, l).

Если узловые значения искомой функции $u_m, m = i, j, k, l$ станут известными, то значения искомой функции $u(x, y, z)$ во внутренних точках конечного элемента Ω^e определяются с помощью следующей функции:

$$\widehat{u}^e(x,y,z) = u_i^e N_i^e(x,y,z) + u_j^e N_j^e(x,y,z) + \\ + u_k^e N_k^e(x,y,z) + u_l^e N_l^e(x,y,z), \quad (8.67)$$

называемой *функцией элемента*.

Слабая формулировка конечно-элементного метода Галеркина, изложенная выше для двумерных задач, полностью сохраняется для трехмерных задач за исключением того, что криволинейные интегралы первого рода заменяются на поверхностные интегралы первого рода, а двойные интегралы по области Ω — на тройные интегралы.

При этом первая формула Грина, используемая в слабой формулировке метода Галеркина, имеет вид

$$\int\limits_{\Omega} u\left[\frac{\partial}{\partial x}\left(\lambda\frac{\partial \mathrm{v}}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda\frac{\partial \mathrm{v}}{\partial y}\right) + \frac{\partial}{\partial z}\left(\lambda\frac{\partial \mathrm{v}}{\partial z}\right)\right]dxdydz = \\ = -\int\limits_{\Omega}\left(\frac{\partial u}{\partial x}\lambda\frac{\partial \mathrm{v}}{\partial x} + \frac{\partial u}{\partial y}\lambda\frac{\partial \mathrm{v}}{\partial y} + \frac{\partial u}{\partial z}\lambda\frac{\partial \mathrm{v}}{\partial z}\right)dxdydz + \iint\limits_{\Gamma} u\cdot\lambda\frac{\partial \mathrm{v}}{\partial n}d\Gamma.$$

В соответствии с количеством нумерованных узлов в конечном элементе Ω^e локальная матрица A^e в системе (8.41) имеет размерность 4×4, причем интегралы в элементах матрицы A^e и компонентах вектора правых частей F^e вычисляются с помощью тех же принципов (теорем о среднем), что и в двумерном случае.

Процедура включения трехмерного элемента Ω^e в ансамбль элементов, локальной матрицы жесткости A^e в глобальную A и локального вектора правых частей F^e в глобальный вектор F остается такой же, как и в двумерном случае.

§ 8.8. Оценка погрешности метода конечных элементов

8.8.1. Погрешность конечно-элементного метода решения задач для обыкновенных дифференциальных уравнений. Границы погрешности конечно-элементного метода оценим вначале на примере первой краевой задачи для обыкновенного дифференциального уравнения второго порядка [29]:

$$Lu = -\frac{d^2u}{dx^2} + a(x)u = f(x); u(0) = u(1) = 0, \tag{8.68}$$

где $a(x)$ — известная положительная функция, удовлетворяющая условию

$$0 < \beta_1 \leqslant a(x) \leqslant \beta_2 < \infty, \quad x \in [0; 1]. \tag{8.69}$$

Пусть конечно-элементное решение имеет вид (M — число конечных элементов)

$$u(x) \approx \widehat{u}(x) = \sum_{m=1}^{M+1} u_m N_m(x). \tag{8.70}$$

Для нахождения погрешности решения (8.70) введем функциональное пространство Соболева $W_p^n[a, b]$ — пространство, для любого элемента $u(x)$ которого выполняется условие

$$\frac{d^\alpha u}{dx^\alpha} \in L_p[a, b] \text{ для всех } \alpha \leqslant n, \tag{8.71}$$

где L_p — пространство функций, интегрируемых со степенью p (интегрируемых по Лебегу). Для задачи (8.68) будем рассматривать пространство Соболева $W_2^1[0; 1]$ функций $u(x)$ со скалярным произведением и нормой вида

$$\begin{gathered}
(u(x), \mathrm{v}(x)) = \int_0^1 \left(u\,\mathrm{v} + \frac{du}{dx}\frac{d\,\mathrm{v}}{dx} \right) dx, \\
(u(x), u(x)) = \|u(x)\|^2_{W_2^1} = \int_0^1 \left[u^2 + \left(\frac{du}{dx} \right)^2 \right] dx.
\end{gathered} \tag{8.72}$$

Здесь нижний индекс $p = 2$ указывает на то, что сама функция $u(x)$ и ее первая производная (верхний индекс $n = 1$) должны быть интегрируемы с квадратом. Таким образом, верхний

индекс 1 определяет наивысший порядок непрерывной производной в подынтегральном выражении.

Например, нормы функций $u_1 = x^2$ и $u_2 = \sin x$ в пространстве Соболева $W_2^1[0;1]$ будут соответственно

$$\|u_1\|_{W_2^1[0;1]} = \left[\int_0^1 \left(x^4 + 4x^2\right) dx\right]^{\frac{1}{2}};$$

$$\|u_2\|_{W_2^1[0;1]} = \left[\int_0^1 \left(\sin^2 x + \cos^2 x\right) dx\right]^{\frac{1}{2}}.$$

Введем далее M-мерное подпространство S^h функций $\mathrm{v}(x) \in \in S^h \subset \overset{\circ}{W}{}_2^1$, где $M = 1/h$, h — шаг разбиения области $x \in [0;1]$, а кружок над символом пространства W означает, что функции $\mathrm{v}(x)$ в граничных точках $x = 0$ и $x = 1$ обращаются в нуль. Рассмотрим класс функций, принадлежащих этому подпространству.

Во-первых, это линейно-непрерывные глобальные базисные функции $N_m(x)$, $m = \overline{2, M}$, определяемые соотношениями (8.3), поскольку эти функции кусочно-непрерывно дифференцируемы, т. е. интеграл (8.72) существует, $N_m(x) \in W_2^1[0;1]$ и на границах

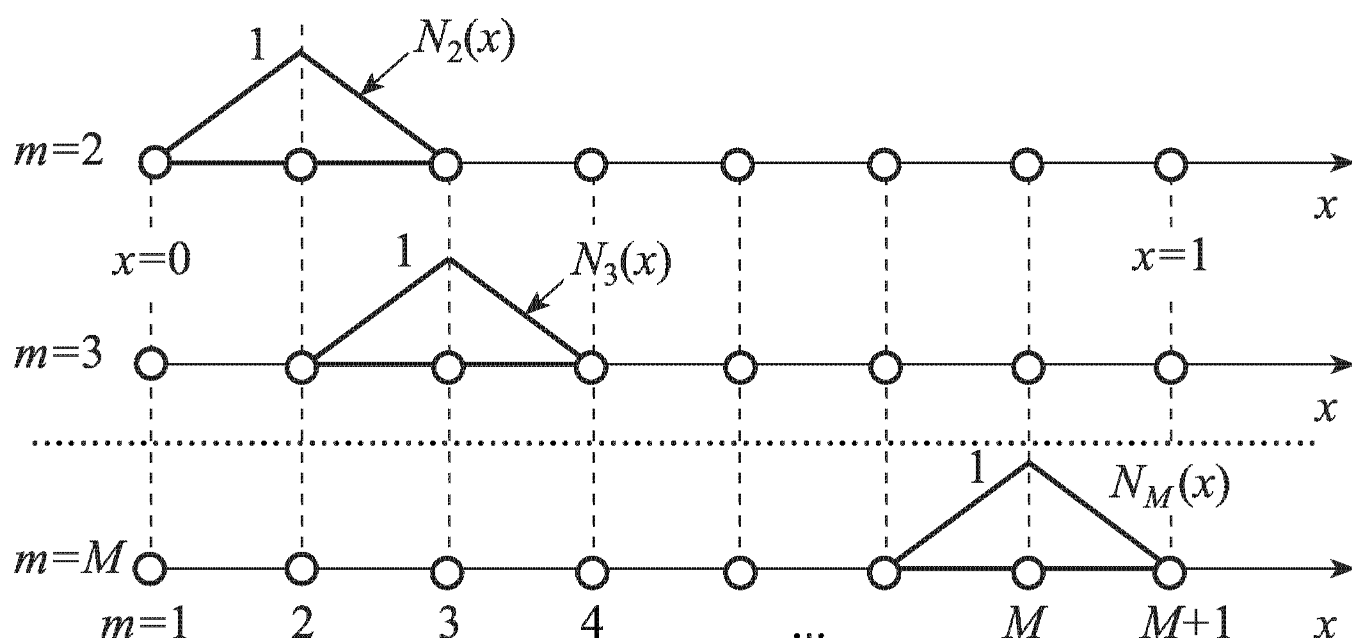


Рис. 8.14. Глобальные базисные функции $N_m(x)$

отрезка $x \in [0;1]$ они принимают нулевые значения (рис. 8.14). При этом базисные функции для граничных узлов пока не учтены.

Во-вторых, это линейная комбинация $\widehat{u}(x)$ базисных функций (8.70), поскольку в силу однородных краевых условий первого рода задачи (8.68) функция $\widehat{u}(x) \in \overset{\circ}{W}{}_2^1$.

В-третьих, функция $\overline{u}(x)$, интерполирующая функцию $u(x)$ из задачи (8.68) и совпадающая с $u(x)$ в нумерованных узлах, то есть $\overline{u}(x_m) = u(x_m)$, $m = \overline{1, M+1}$, поскольку $\overline{u}(x)$ на границах отрезка $x \in [0;1]$ также равна нулю и имеет кусочно-непрерывные производные первого порядка.

На основе этого можно заключить, что и разность $\mathrm{v}(x) = \overline{u}(x) - \widehat{u}(x)$ принадлежит подпространству $S^h \subset \overset{\circ}{W}{}_2^1$. Таким образом, к функциям $\mathrm{v}(x) \in \overset{\circ}{W}{}_2^1$ можно отнести множество функций $\{N_m(x), \widehat{u}(x), \overline{u}(x), \overline{u} - \widehat{u}\}$.

Пространство $\overset{\circ}{W}{}_2^1$ интересно тем, что если в качестве контрольной функции взять элемент этого пространства $\mathrm{v}\ (x) \in \overset{\circ}{W}{}_2^1$, например базисную функцию $N_m(x)$, то скалярные произведения, записанные на основе (8.68) и (8.70), будут иметь вид

$$(Lu, \mathrm{v}) = (f, \mathrm{v})\,, \tag{8.73}$$

$$\big(L\widehat{u}, \mathrm{v}\big) = (f, \mathrm{v}) + (R, \mathrm{v})\,, \tag{8.74}$$

где $R(x)$ — невязка, являющаяся результатом подстановки (8.70) в (8.68). Для $\mathrm{v} = N_m(x)$ скалярное произведение $(R, \mathrm{v}) \equiv 0$ в силу метода Галеркина взвешенных невязок и ортогональности $R(x)$ и $N_m(x)$.

Тогда, вычитая (8.74) из (8.73) (при $(R, \mathrm{v}) \equiv 0$), получим

$$(L\widehat{e}, \mathrm{v}) = 0, \tag{8.75}$$

$\widehat{e} = u - \widehat{u}$, причем можно показать, что (8.75) имеет место для любой функции $\mathrm{v}(x)$ из перечисленного множества функций пространства $\overset{\circ}{W}{}_2^1$.

В методе конечных элементов оценку погрешности будем сравнивать с оценкой погрешности известных методов, например метода интерполяции.

Тогда в соответствии с линейным свойством скалярного произведения имеем следующее равенство:

$$(L\widehat{e}, \widehat{e}) = (L\widehat{e}, \widehat{e} + \overline{u} - \overline{u}) = (L\widehat{e}, u - \overline{u}) + (L\widehat{e}, \overline{u} - \widehat{u})\,. \tag{8.76}$$

Поскольку разность $\overline{u}-\widehat{u}$ принадлежит подпространству $S^h \subset \overset{\circ}{W}{}_2^1$, то в соответствии с (8.75) второе слагаемое в правой части выражения (8.76) равно нулю и оно принимает вид

$$(L\widehat{e},\widehat{e}) = (L\widehat{e}, u-\overline{u}),$$

или

$$\int\limits_0^1 \left(-\frac{d^2\widehat{e}}{dx^2}+a(x)\widehat{e}\right)\widehat{e}dx = \int\limits_0^1 \left(-\frac{d^2\widehat{e}}{dx^2}+a(x)\widehat{e}\right)(u-\overline{u})\,dx. \quad (8.77)$$

После интегрирования по частям первых слагаемых в подынтегральных выражениях равенство (8.77) приводится к виду (с учетом однородных граничных условий задачи (8.68))

$$\int\limits_0^1 \left[\left(\frac{d\widehat{e}}{dx}\right)^2 + a\widehat{e}^2\right]dx = \int\limits_0^1 \left[\frac{d\widehat{e}}{dx}\frac{d}{dx}(u-\overline{u}) + a(x)\widehat{e}\,(u-\overline{u})\right]dx. \quad (8.78)$$

Обозначим левую часть в (8.78) через I, тогда на основе неравенства Шварца

$$\int\limits_a^b u\cdot v dx \leqslant \left(\int\limits_a^b u^2 dx\right)^{\frac{1}{2}} \cdot \left(\int\limits_a^b v^2 dx\right)^{\frac{1}{2}}$$

для любых u и v имеем

$$I \leqslant \left[\int\limits_0^1 \left(\frac{d\widehat{e}}{dx}\right)^2 dx\right]^{\frac{1}{2}} \cdot \left\{\int\limits_0^1 \left[\frac{d}{dx}(u-\overline{u})\right]^2 dx\right\}^{\frac{1}{2}} + \\ + a_{\max}\left(\int\limits_0^1 \widehat{e}^2 dx\right)^{\frac{1}{2}} \cdot \left[\int\limits_0^1 (u-\overline{u})^2\,dx\right]^{\frac{1}{2}}. \quad (8.79)$$

Из определения соболевской нормы (8.72) следует, что

$$\left[\int\limits_0^1 \left(\frac{d\widehat{e}}{dx}\right)^2 dx\right]^{\frac{1}{2}} / \, \|\widehat{e}\|_{W_2^1} \leqslant 1, \; \left(\int\limits_0^1 \widehat{e}^2 dx\right)^{\frac{1}{2}} / \, \|\widehat{e}\|_{W_2^1} \leqslant 1.$$

Тогда из (8.79) находим

$$I \leqslant \|\widehat{e}\|_{W_2^1} \times$$
$$\times \left\{ \left[\int\limits_0^1 \left[\frac{d}{dx}(u-\overline{u})\right]^2 dx\right]^{\frac{1}{2}} + a_{\max}\left[\int\limits_0^1 (u-\overline{u})^2\, dx\right]^{\frac{1}{2}} \right\} \leqslant$$
$$\leqslant \max(1,\beta_2)\, \|\widehat{e}\|_{W_2^1} \left(\left\{ \int\limits_0^1 \left[\frac{d}{dx}(u-\overline{u})\right]^2 dx \right\}^{\frac{1}{2}} + \right.$$
$$\left. + \left[\int\limits_0^1 (u-\overline{u})^2\, dx\right]^{\frac{1}{2}} \right) \leqslant \max(1,\beta_2)\, \|\widehat{e}\|_{W_2^1}\, \|u-\overline{u}\|_{W_2^1}\,. \quad (8.80)$$

Здесь β_2 — верхняя граница коэффициента $a(x)$ в ограничениях (8.69). Третье неравенство в цепочке (8.80) основано на известной теореме о том, что среднее арифметическое двух неотрицательных выражений ($a_1 \geqslant 0$, $a_2 \geqslant 0$) не ниже их среднего геометрического:

$$(a_1+a_2)/2 \geqslant \sqrt{a_1 \cdot a_2}\,. \quad (8.81)$$

На основе этого неравенства требуется доказать, что

$$a_1^{1/2} + a_2^{1/2} \leqslant \sqrt{2}\,(a_1+a_2)^{1/2}. \quad (8.82)$$

Действительно, прибавим к левой и правой частям (8.81) сумму (a_1+a_2), получим

$$2\cdot(a_1+a_2) \geqslant 2\sqrt{a_1 a_2} + a_1 + a_2 = (a_1^{1/2} + a_2^{1/2})^2,$$

откуда сразу следует (8.82) и третье неравенство в (8.80). Таким образом, из (8.80) получаем верхнюю оценку для выражения I:

$$I = \int_0^1 \left[\left(\frac{d\widehat{e}}{dx^2} + a\widehat{e}^2\right)\right] dx \leqslant \max(1, \beta_2)\sqrt{2}\, \|\widehat{e}\|_{W_2^1} \|u - \overline{u}\|_{W_2^1}\,. \tag{8.83}$$

Для нижней оценки оператора $I = (L\widehat{e}, \widehat{e})$ заметим, что

$$I = \int_0^1 \left[\left(\left(\frac{d\widehat{e}}{dx}\right)^2 + a\widehat{e}^2\right)\right] dx \geqslant$$

$$\geqslant \min(1, \beta_1)\cdot \int_0^1 \left[\left(\left(\frac{d\widehat{e}}{dx}\right)^2 + \widehat{e}^2\right)\right] dx =$$

$$= \min(1, \beta_1)\cdot \|\widehat{e}\|^2_{W_2^1}\,, \tag{8.84}$$

где β_1 — нижняя грань коэффициента $a(x)$ в (8.69).

Из сравнения оценок (8.83) и (8.84) следует, что

$$\min(1, \beta_1)\cdot \|\widehat{e}\|^2_{W_2^1} \leqslant \max(1, \beta_2)\sqrt{2}\, \|\widehat{e}\|_{W_2^1} \|u - \overline{u}\|_{W_2^1}\,,$$

откуда

$$\|\widehat{e}\| = \|u - \widehat{u}\| \leqslant c\,\|u - \overline{u}\|\,, \quad c = \sqrt{2}\,\max(1, \beta_2)\,/\,\min(1, \beta_1)\,, \tag{8.85}$$

т. е. норма погрешности конечно-элементного решения (8.70) не превышает нормы погрешности *интерполяционной* функции.

Наконец, свяжем норму погрешности конечно-элементного решения $\widehat{u}$ с размером h конечного элемента. Используем погрешность линейной интерполяции по Лагранжу $\overline{u}(x)$ в точке x на отрезке $[x_{j-1}, x_j]$ (решение на каждом конечном элементе с помощью базисных функций (8.3) является линейной функцией):

$$|u - \overline{u}(x)| = \left|\frac{u''(\xi)}{2!}(x - x_{j-1})(x - x_j)\right| \leqslant \max_{x\in[x_{j-1},x_j]} \left|u''(x)\right| \frac{h^2}{2} =$$

$$= c_{1j}h^2, \xi \in (x_{j-1}, x_j)\,, \tag{8.86}$$

$$\left|\frac{d}{dx}\,(u - \overline{u}(x))\right| = \left|\frac{u''(\xi)}{2!}\,[(x - x_j) + (x - x_{j-1})]\,\right| \leqslant$$

$$\leqslant \max_{x\in[x_{j-1},x_j]} \left|u''(x)\right| 2h/2 \leqslant c_{2j}h. \quad (8.87)$$

Интегрируя погрешности (8.86) и (8.87) по всему отрезку $x \in \in [0;\,1]$ и выбирая максимальные на этом отрезке константы, получаем

$$\int_0^1 (u-\overline{u})^2\,dx \leqslant c_1 h^4; \quad \int_0^1 \left[\frac{d}{dx}(u-\overline{u})\right]^2 dx \leqslant c_2 h^2.$$

Тогда соболевская норма погрешности интерполяционной формулы оценивается выражением

$$\|u-\overline{u}\|_{W_2^1} = \left\{\int_0^1 \left((u-\overline{u})^2 + \left[\frac{d}{dx}(u-\overline{u})\right]^2\right) dx\right\}^{1/2} \leqslant$$

$$\leqslant \left(c_1 h^4 + c_2 h^2\right)^{1/2} \leqslant c_3 h. \quad (8.88)$$

На основе (8.85) и (8.88) заключаем, что погрешность конечно-элементного решения (8.70) оценивается сверху выражением

$$\|u-\widehat{u}\| \leqslant c_3\sqrt{2}\cdot[\max(1,\beta_2)\,/\min(1,\beta_1)]\cdot h, \quad (8.89)$$

т. е. имеет место первый порядок по шагу, что согласуется с линейностью конечных элементов для задачи (8.68).

8.8.2. Погрешность конечно-элементного метода решения задач для уравнений в частных производных. Изложенный выше метод оценки погрешности конечно-элементного метода решения задач для обыкновенных дифференциальных уравнений легко распространяется на уравнения в частных производных.

Погрешность конечно-элементного метода для задач математической физики рассмотрим на примере следующей задачи Дирихле для уравнения Пуассона в плоской области $\Omega+\Gamma$:

$$Lu = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + a(x,y)u = q(x,y),\,(x,y)\in\Omega, \quad (8.90)$$

$$u(x,y) = 0, \quad (x,y)\in\Gamma, \quad (8.91)$$

в которой $a(x, y)$ — положительная функция, удовлетворяющая условиям

$$0 < \gamma_1 \leqslant a(x, y) \leqslant \gamma_2 < \infty, \tag{8.92}$$

а оператор Lu положительно определен.

Для задачи (8.90), (8.91) введем пространство Соболева W_2^1 функций $u(x, y)$ со скалярным произведением

$$(u(x, y), \mathrm{v}(x, y)) = \iint\limits_{\Omega} \left(u\mathrm{v} + \frac{\partial u}{\partial x}\frac{\partial \mathrm{v}}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \mathrm{v}}{\partial y} \right) dxdy$$

и нормой

$$\|u(x, y)\|^2_{W_2^1} = \\ = (u(x, y), u(x, y)) = \iint\limits_{\Omega} \left[u^2 + \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \right] dxdy.$$

Повторяя выкладки, аналогичные выкладкам для обыкновенных дифференциальных уравнений, вплоть до соотношения (8.77), находим

$$\iint\limits_{\Omega} \left[-\left(\frac{\partial^2 \widehat{e}}{\partial x^2} + \frac{\partial^2 \widehat{e}}{\partial y^2} \right) + a(x, y)\widehat{e} \right] \cdot \widehat{e} dxdy = \\ = \iint\limits_{\Omega} \left[-\left(\frac{\partial^2 \widehat{e}}{\partial x^2} + \frac{\partial^2 \widehat{e}}{\partial y^2} \right) + a(x, y)\widehat{e} \right] \cdot (u - \overline{u})\, dxdy. \tag{8.93}$$

Применим к первым двум слагаемым подынтегральных выражений левой и правой частей равенства (8.93) первую формулу Грина, в которой криволинейный интеграл по границе Г равен нулю, поскольку $u|_\Gamma \equiv 0$; $\widehat{u}|_\Gamma \equiv 0$; $\widehat{e} = u - \widehat{u} \equiv 0$, получим

$$\iint\limits_{\Omega} \left[\left(\frac{\partial \widehat{e}}{\partial x}\right)^2 + \left(\frac{\partial \widehat{e}}{\partial y}\right)^2 + a\widehat{e}^2 \right] dxdy = \\ = \iint\limits_{\Omega} \left[\frac{\partial \widehat{e}}{\partial x}\frac{\partial}{\partial x}(u - \overline{u}) + \frac{\partial \widehat{e}}{\partial y}\frac{\partial}{\partial y}(u - \overline{u}) + a\widehat{e}(u - \overline{u}) \right] dxdy.$$

Правую часть этого равенства оценим сверху с помощью неравенства Шварца, обозначив выражение в левой части через I:

$$I \leqslant \left[\iint\limits_{\Omega} \left(\frac{\partial \widehat{e}}{\partial x}\right)^2 dxdy\right]^{\frac{1}{2}} \cdot \left\{\iint\limits_{\Omega} \left[\frac{\partial}{\partial x}(u - \overline{u})\right]^2 dxdy\right\}^{\frac{1}{2}} + \\ + \left[\iint\limits_{\Omega} \left(\frac{\partial \widehat{e}}{\partial y}\right)^2 dxdy\right]^{\frac{1}{2}} \cdot \left\{\iint\limits_{\Omega} \left[\frac{\partial}{\partial y}(u - \overline{u})\right]^2 dxdy\right\}^{\frac{1}{2}} + \\ + \gamma_2 \left(\iint\limits_{\Omega} \widehat{e}^2 dxdy\right)^{\frac{1}{2}} \cdot \left[\iint\limits_{\Omega} (u - \overline{u})^2 dxdy\right]^{\frac{1}{2}}. \quad (8.94)$$

В соответствии с соболевской нормой имеем неравенства

$$\left[\iint\limits_{\Omega} \left(\frac{\partial \widehat{e}}{\partial x}\right)^2 dxdy\right]^{\frac{1}{2}} / \, \|\widehat{e}\|_{W_2^1} \leqslant 1,$$

$$\left[\iint\limits_{\Omega} \left(\frac{\partial \widehat{e}}{\partial y}\right)^2 dxdy\right]^{\frac{1}{2}} / \, \|\widehat{e}\|_{W_2^1} \leqslant 1,$$

$$\left(\iint\limits_{\Omega} \widehat{e}^2 dxdy\right)^{\frac{1}{2}} / \, \|\widehat{e}\|_{W_2^1} \leqslant 1.$$

Используя эти неравенства, из (8.94) получаем

$$I \leqslant \max(1, \gamma_2) \, \|\widehat{e}\|_{W_2^1} \cdot \left(\left\{\iint\limits_{\Omega} \left[\frac{\partial}{\partial x}(u - \overline{u})\right]^2 dxdy\right\}^{\frac{1}{2}} + \right. \\ \left. + \left\{\iint\limits_{\Omega} \left[\frac{\partial}{\partial y}(u - \overline{u})\right]^2 dxdy\right\}^{\frac{1}{2}} + \left[\iint\limits_{\Omega} (u - \overline{u})^2 dxdy\right]^{\frac{1}{2}}\right). \quad (8.95)$$

Чтобы связать выражение в круглых скобках неравенства (8.95) с соболевской нормой, воспользуемся неравенством

$$a^{1/2} + b^{1/2} + c^{1/2} \leqslant \sqrt{3}\,(a + b + c)^{1/2},$$

справедливым для любых неотрицательных a, b, c.

Используя это неравенство, из (8.95) получаем оценку

$$I \leqslant \sqrt{3}\, \max{(1,\gamma_2)}\, \|\widehat{e}\|_{W_2^1} \left(\iint\limits_{\Omega} \left\{ \left[\frac{\partial}{\partial x}(u-\overline{u})\right]^2 + \left[\frac{\partial}{\partial y}(u-\overline{u})\right]^2 + \right.\right.$$

$$\left.\left. + (u-\overline{u})^2 \right\} dxdy \right)^{\frac{1}{2}} = \sqrt{3}\, \max{(1,\gamma_2)}\, \|\widehat{e}\|_{W_2^1}\, \|u-\overline{u}\|_{W_2^1}$$

или

$$\iint\limits_{\Omega} \left[\left(\frac{\partial \widehat{e}}{\partial x}\right)^2 + \left(\frac{\partial \widehat{e}}{\partial y}\right)^2 + a(x,y)\widehat{e}^2 \right] dxdy \leqslant$$

$$\leqslant \sqrt{3}\, \max{(1,\gamma_2)}\, \|\widehat{e}\|_{W_2^1}\, \|u-\overline{u}\|_{W_2^1}\,. \quad (8.96)$$

Используя положительную определенность оператора Lu, находим нижнюю границу интеграла в левой части выражения (8.96):

$$(L\widehat{e},\widehat{e}) = \iint\limits_{\Omega} \left[\left(\frac{\partial \widehat{e}}{\partial x}\right)^2 + \left(\frac{\partial \widehat{e}}{\partial y}\right)^2 + a(x,y)\widehat{e}^2 \right] dxdy \geqslant$$

$$\geqslant \min{(1,\gamma_1)}\, \|\widehat{e}\|^2_{W_2^1}\,. \quad (8.97)$$

Сравнение неравенств (8.96) и (8.97) приводит к оценке для нормы погрешности $\|u-\overline{u}\|$ в пространстве Соболева:

$$\|\widehat{e}\|_{W_2^1} \leqslant D\, \|u-\overline{u}\|_{W_2^1}\,, \quad D = \sqrt{3}\, \max{(1,\gamma_2)} \,/\, \min{(1,\gamma_1)}\,, \quad (8.98)$$

из которой следует, что и для уравнений в частных производных погрешность конечно-элементного метода не превышает погрешности *интерполяции* с точностью до константы D.

§ 8.9. Вариационный принцип в МКЭ

При решении краевых задач для дифференциальных уравнений с помощью вариационного метода необходимо строить такой функционал, минимум которого достигается на допустимых функциях, удовлетворяющих уравнению Эйлера для этого

функционала, причем уравнение Эйлера совпадает с искомым дифференциальным уравнением. То есть минимум упомянутого функционала достигается на функциях, удовлетворяющих решению исходной дифференциальной задачи.

8.9.1. Введение в вариационное исчисление. Вариационное исчисление связано с отысканием стационарных значений функционалов [26,27], представляющих собой определенные интегралы от специальным образом построенных функций и принимающих числовое значение при подстановке каждой конкретной функции в подынтегральное выражение.

Основная задача вариационного исчисления состоит в отыскании такой функции $F(x)$, чтобы при произвольном бесконечно малом изменении этой функции $\delta F(x)$ величина функционала (определенного интеграла от этой функции) оставалась неизменной, т. е. функционал принимал стационарное значение.

Рассмотрим функционал

$$I = \int_a^b F(x, \varphi, \varphi_x) dx, \tag{8.99}$$

где x — независимая переменная; $\varphi(x)$ — функция этой переменной, $\varphi_x = d\varphi/dx$. Варьирование функционала I вызывается бесконечно малым изменением (варьированием) функции $F(x)$:

$$\delta I = \int_a^b \delta F(x) dx = \int_a^b \left(\frac{\partial F}{\partial \varphi} \delta\varphi + \frac{\partial F}{\partial \varphi_x} \delta\varphi_x \right) dx. \tag{8.100}$$

Чтобы вынести за скобку подынтегрального выражения в (8.100) вариацию $\delta\varphi$, воспользуемся равенством

$$\delta\varphi_x = \delta\left(\frac{d\varphi}{dx}\right) = \frac{d}{dx}(\delta\varphi), \tag{8.101}$$

после чего, проинтегрировав второе слагаемое в (8.100) по частям, получим

$$\delta I = \int_a^b \left[\frac{\partial F}{\partial \varphi} - \frac{d}{dx}\left(\frac{\partial F}{\partial \varphi_x}\right)\right] \delta\varphi dx + \frac{\partial F}{\partial \varphi_x}\delta\varphi \Big|_a^b. \tag{8.102}$$

Функционал (8.99) принимает стационарное значение, если $\delta I = 0$, откуда, в силу произвольности $\delta\varphi$, имеем

$$\frac{\partial F}{\partial \varphi} - \frac{d}{dx}\frac{\partial F}{\partial \varphi_x} = 0; \tag{8.103}$$

$$\frac{\partial F\,(a)}{\partial \varphi_x} = \frac{\partial F\,(b)}{\partial \varphi_x} = 0 \tag{8.104}$$

или

$$\delta\varphi\,(a) = \delta\varphi\,(b) = 0, \tag{8.105}$$

что соответствует постоянству функции φ на границах $x = a$, $x = b$, т. е. $\varphi\,(a) = \mathrm{const}_1$, $\varphi\,(b) = \mathrm{const}_2$.

Рассмотрим теперь функционал от функции трех независимых переменных в области V, ограниченной границей Г:

$$I = \int\limits_V F\,(x, y, z, \varphi, \varphi_x, \varphi_y, \varphi_z) dV. \tag{8.106}$$

Произвольному бесконечно малому изменению функции $F\,(x, y, z)$ соответствует вариация функционала

$$\delta I = \int\limits_V \left(\frac{\partial F}{\partial \varphi}\delta\varphi + \frac{\partial F}{\partial \varphi_x}\delta\varphi_x + \frac{\partial F}{\partial \varphi_y}\delta\varphi_y + \frac{\partial F}{\partial \varphi_z}\delta\varphi_z \right) dV. \tag{8.107}$$

Используя соотношение (8.101), получаем выражение

$$\delta I = \int\limits_V \left[\frac{\partial F}{\partial \varphi}\delta\varphi + \frac{\partial F}{\partial \varphi_x}\frac{\partial}{\partial x}\,(\delta\varphi) + \frac{\partial F}{\partial \varphi_y}\frac{\partial}{\partial y}\,(\delta\varphi) + \frac{\partial F}{\partial \varphi_z}\frac{\partial}{\partial z}\,(\delta\varphi) \right] dV. \tag{8.108}$$

Применяя ко второму слагаемому формулу дифференцирования произведения, а затем формулу Остроградского–Гаусса, находим

$$\int\limits_V \frac{\partial F}{\partial \varphi_x}\frac{\partial}{\partial x}\,(\delta\varphi)\,dV = \int\limits_V \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial \varphi_x}\delta\varphi \right) dV - \int\limits_V \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial x} \right) \delta\varphi dV =$$

$$= \int\limits_\Gamma n_x \frac{\partial F}{\partial \varphi_x}\delta\varphi\, d\Gamma - \int\limits_V \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial \varphi_x} \right) \delta\varphi dV, \tag{8.109}$$

где n_x — первый направляющий косинус нормали к границе Г.

Таким образом, из (8.108) и (8.109) имеем следующее выражение:

$$\delta I = \int\limits_V \left[\frac{\partial F}{\partial \varphi} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial \varphi_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial \varphi_y}\right) - \frac{\partial}{\partial z}\left(\frac{\partial F}{\partial \varphi_z}\right)\right] \delta\varphi dV +$$

$$+ \int\limits_\Gamma \left[n_x \frac{\partial F}{\partial \varphi_x} + n_y \frac{\partial F}{\partial \varphi_y} + n_z \frac{\partial F}{\partial \varphi_z}\right] \delta\varphi d\Gamma. \quad (8.110)$$

Стационарное значение функционала (8.106) достигается при равенстве нулю подынтегральных выражений в квадратных скобках равенства (8.110), т. е. получаем уравнение Эйлера

$$\frac{\partial F}{\partial \varphi} - \left[\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial \varphi_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial \varphi_y}\right) + \frac{\partial}{\partial z}\left(\frac{\partial F}{\partial \varphi_z}\right)\right] = 0 \quad (8.111)$$

с ограничением

$$n_x \frac{\partial F}{\partial \varphi_x} + n_y \frac{\partial F}{\partial \varphi_y} + n_z \frac{\partial F}{\partial \varphi_z} = 0. \quad (8.112)$$

Соотношения (8.110)–(8.112) соответствуют вариационной формулировке задач переноса тепла и массы.

Действительно, рассмотрим функционал

$$I = \int\limits_V F(\varphi, \varphi_x, \varphi_y, \varphi_z) dV =$$

$$= \frac{1}{2} \int\limits_V \left[\lambda_{xx}\left(\frac{\partial \varphi}{\partial x}\right)^2 + \lambda_{yy}\left(\frac{\partial \varphi}{\partial y}\right)^2 + \lambda_{zz}\left(\frac{\partial \varphi}{\partial z}\right)^2 - 2Q\varphi\right] dV. \quad (8.113)$$

Тогда для уравнения Эйлера (8.111) и ограничений (8.112) имеем

$$\frac{\partial}{\partial x}\left(\lambda_{xx}\frac{\partial \varphi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda_{yy}\frac{\partial \varphi}{\partial y}\right) + \frac{\partial}{\partial z}\left(\lambda_{zz}\frac{\partial \varphi}{\partial z}\right) + Q = 0, \quad (8.114)$$

$$\lambda_{xx}\frac{\partial \varphi}{\partial x} n_x + \lambda_{yy}\frac{\partial \varphi}{\partial y} n_y + \lambda_{zz}\frac{\partial \varphi}{\partial z} n_z = 0, \quad (8.115)$$

т. е. получаем стационарное уравнение теплопроводности с источниками Q в ортотропной среде с главными компонентами тензора теплопроводности λ_{xx}, λ_{yy}, λ_{zz} и граничное условие второго рода на границе Γ, ограничивающей тело V.

Граничное условие (8.115) автоматически удовлетворяется при задании функционала в форме (8.113) и поэтому называется естественным.

8.9.2. Конечно-элементный вариационный принцип на основе симметричного дифференциального оператора. Вариационный метод Релея–Ритца. Если задан функционал (вариационный принцип), то, как правило, можно найти уравнение Эйлера. К сожалению, для дифференциальной задачи не всегда можно построить функционал. Однако в одном важном случае, а именно в случае симметричного дифференциального оператора, это можно сделать.

Пусть дана дифференциальная задача

$$L\varphi + p = 0 \text{ в } \Omega; \tag{8.116}$$

$$B\varphi + r = 0 \text{ на } \Gamma, \tag{8.117}$$

где L и B — линейные дифференциальные операторы, а p и r — известные функции.

Пусть множество Φ функций $\theta \in \Phi$ непрерывно дифференцируемых в области $\Omega + \Gamma$, удовлетворяет на границе Γ однородному краевому условию (8.117), т. е.

$$B\theta = 0 \text{ на } \Gamma. \tag{8.118}$$

Тогда оператор L называется *симметричным* относительно множества Φ, если для двух элементов θ и ϑ этого множества скалярное произведение удовлетворяет равенству

$$(\theta,\ L\vartheta) = (\vartheta, L\theta)$$

или

$$\int\limits_{\Omega} \theta L\vartheta\, d\Omega = \int\limits_{\Omega} \vartheta L\theta\, d\Omega, \tag{8.119}$$

т. е. оператор L — *самосопряженный* в Ω с однородным граничным условием на Γ.

Симметричный оператор L называется *положительно определенным* относительно этого множества функций Φ, если для

любого элемента θ из данного множества

$$(\theta, L\theta) = \int_{\Omega} \theta \, L\theta \; d\Omega \geqslant 0, \tag{8.120}$$

где равенство имеет место в том и только в том случае, когда $\theta \equiv 0$ в Ω.

Например, оператор $L = -d^2/dx^2$ на отрезке $x \in [0;\ 1]$ и множестве функций, удовлетворяющих условиям $\theta(0) = \theta(1) = 0$, является симметричным и положительно определенным. Действительно, для двух элементов θ и ϑ после двойного интегрирования по частям будем иметь

$$\int_0^1 \left(-\theta \frac{d^2\vartheta}{dx^2}\right) dx = \left(-\theta \frac{d\vartheta}{dx}\right)_0^1 + \int_0^1 \frac{d\theta}{dx}\frac{d\vartheta}{dx} dx =$$

$$= 0 + \left(\vartheta \frac{d\theta}{dx}\right)_0^1 + \int_0^1 \left(-\vartheta \frac{d^2\theta}{dx^2}\right) dx. \tag{8.121}$$

Так как

$$\theta \left.\frac{d\vartheta}{dx}\right|_0^1 = \vartheta \left.\frac{d\theta}{dx}\right|_0^1 = 0,$$

то $(\theta, L\vartheta) = (\vartheta, L\theta)$, что и требовалось доказать.

Кроме того, заменяя в первом равенстве ϑ на θ, получаем

$$\int_0^1 \left(-\theta \frac{d^2\theta}{dx^2}\right) dx = \int_0^1 \left(\frac{d\theta}{dx}\right)^2 dx \geqslant 0, \tag{8.122}$$

причем равенство нулю в (8.123) имеет место тогда и только тогда, когда $d\theta/dx = 0$, а это при однородных граничных условиях $\theta(0) = \theta(1) = 0$ выполняется при $\theta \equiv 0$.

Таким образом, оператор $L = -d^2/dx^2 \geqslant 0$.

Пусть дана задача (8.116), (8.117) и L — симметричный оператор относительно множества $\varphi \in \Phi$, удовлетворяющий однородному краевому условию (8.118), а ψ — некоторая функция, для которой на границе Γ выполняется неоднородное краевое

условие

$$B\psi + r = 0 \text{ на } \Gamma.$$

Тогда функционал

$$I = \int\limits_{\Omega} (\varphi - \psi) \left[\frac{1}{2} L(\varphi - \psi) + L\psi + p\right] d\Omega \tag{8.123}$$

принимает стационарное значение на решении φ краевой задачи (8.116), (8.117).

Действительно, вариация функции φ (т. е. переход от φ к $\varphi + \delta\varphi$) определяет вариацию функционала (8.123):

$$\delta I = \int\limits_{\Omega} \left\{\delta\varphi \left[\frac{1}{2} L(\varphi - \psi) + L\psi + p\right] + \frac{1}{2}(\varphi - \psi) L(\delta\varphi)\right\} d\Omega. \tag{8.124}$$

В силу симметричности оператора L на множестве функций, удовлетворяющих (8.118), выполняется равенство

$$\int\limits_{\Omega} (\varphi - \psi) L(\delta\varphi)\, d\Omega = \int\limits_{\Omega} \delta\varphi\, L(\varphi - \psi)\, d\Omega. \tag{8.125}$$

Тогда из (8.124) имеем

$$\delta I = \int\limits_{\Omega} \left\{\delta\varphi \left[\frac{1}{2} L(\varphi - \psi) + L\psi + p\right] + \frac{1}{2}\delta\varphi L(\varphi - \psi)\right\} d\Omega =$$

$$= \int\limits_{\Omega} \delta\varphi \left[L(\varphi - \psi) + L\psi + p\right] d\Omega = \int\limits_{\Omega} \delta\varphi (L\varphi + p)\, d\Omega, \tag{8.126}$$

и так как $\delta\varphi$ произвольно, то для стационарности функционала (8.123) (т. е. при $\delta I = 0$) выполняется равенство

$$L\varphi + p = 0 \text{ на } \Omega, \tag{8.127}$$

т. е. получается исходное дифференциальное уравнение (8.116).

На основе понятия симметричного оператора L рассмотрим *вариационный метод Релея–Ритца* приближенного решения задачи (8.116), (8.117), в которой оператор L — симметричен относительно множества функций, удовлетворяющих краевому условию (8.118). Для этого выберем функцию ψ, удовлетворяющую

неоднородным краевым условиям (8.117) задачи, т. е.

$$B\psi + r = 0 \text{ на } \Gamma, \tag{8.128}$$

и определим систему базисных функций N_m, $m = \overline{1, M}$, такую что

$$BN_m = 0 \text{ на } \Gamma. \tag{8.129}$$

Тогда аппроксимация для функции φ имеет вид

$$\varphi \approx \widehat{\varphi} = \psi + \sum_{m=1}^{M} a_m \cdot N_m, \tag{8.130}$$

а краевые условия (8.117) на Г выполняются автоматически для всех значений постоянных a_1, a_2, ..., a_M. Метод Релея–Ритца состоит в получении для функционала

$$I = \int_{\Omega} \left(\widehat{\varphi} - \psi\right) \left[\frac{1}{2} L \left(\widehat{\varphi} - \psi\right) + L\psi + p\right] d\Omega \tag{8.131}$$

стационарного значения относительно параметров a_1, a_2, ..., a_M.

Подставляя (8.130) в (8.131), находим

$$I = \sum_{l=1}^{M} \sum_{m=1}^{M} (a_l a_m/2) \int_{\Omega} N_l L N_m d\Omega + \\ + \sum_{l=1}^{M} a_l \int_{\Omega} N_l \left(L\psi + p\right) d\Omega. \tag{8.132}$$

Выражение (8.132) принимает стационарное значение, если

$$\frac{\partial I}{\partial a_1} = \frac{\partial I}{\partial a_2} = \ldots = \frac{\partial I}{\partial a_M} = 0. \tag{8.133}$$

Вычислив эти производные от (8.132), приходим к системе линейных уравнений относительно a_1, a_2, ..., a_M:

$$\sum_{m=1}^{M} a_m \int_{\Omega} N_l L N_m \, d\Omega = -\int_{\Omega} N_l \left(L\psi + p\right) \, d\Omega, \\ l = \overline{1, \, M}. \tag{8.134}$$

Эта система в матричной форме записывается так:

$$Ka = f, \tag{8.135}$$

где K — симметрическая матрица.

Компоненты матрицы K и вектора f вычисляются следующим образом:

$$k_{lm} = \int\limits_{\Omega} N_l L N_m d\Omega; \quad f_l = -\int\limits_{\Omega} N_l \left(L\psi + p\right) d\Omega. \tag{8.136}$$

В системе (8.134) весовые функции N_l совпали с базисными, что тождественно методу Галеркина. Таким образом, методы Релея–Ритца и Галеркина совпадают для симметричных операторов L. Разумеется, метод Галеркина справедлив для всех операторов и в этом смысле является общим.

8.9.3. Решение задач с помощью конечно-элементного вариационного принципа. Рассмотрим конечно-элементный вариационный метод на примере решения задачи

$$\frac{d^2\varphi}{dx^2} - \varphi = 0, \quad x \in (0;1)\,; \tag{8.137}$$

$$\frac{d\varphi\,(0)}{dx} + q = 0, \quad x = 0; \tag{8.138}$$

$$\frac{d\varphi\,(1)}{dx} + \alpha\left(\varphi\,(1) - \beta\right) = 0, \quad x = 1. \tag{8.139}$$

Составим следующий функционал, удовлетворяющий этой задаче:

$$I = \int\limits_0^1 \frac{1}{2}\left[\left(\frac{d\varphi}{dx}\right)^2 + \varphi^2\right] dx - \left[q\varphi - \frac{\alpha}{2}\left(\varphi - \beta\right)^2\right]. \tag{8.140}$$

Действительно, вариация функционала (8.140) приводит к соотношению (с учетом интегрирования по частям)

$$\delta I = \int\limits_0^1 \left(-\frac{d^2\varphi}{dx^2} + \varphi\right)\delta\varphi\, dx + \frac{d\varphi}{dx}\delta\varphi\bigg|_0^1 - \left[q - \alpha\left(\varphi - \beta\right)\right]\delta\varphi,$$

причем стационарное значение функционала достигается на функциях φ, удовлетворяющих задаче (1.137)–(1.139). Представим решение в виде линейной комбинации кусочно-линейных базисных функций N_m, $m = \overline{1, M+1}$ (8.3),

$$\varphi \approx \widehat{\varphi} = \sum_{m=1}^{M+1} \varphi_m N_m \tag{8.141}$$

и, подставив его в функционал (8.140), получим

$$I(\varphi_1, \ldots, \varphi_{M+1}) = \\ = \int_0^1 \frac{1}{2}\left[\left(\frac{d}{dx}\sum_{m=1}^{M+1}\varphi_m N_m\right)^2 + \left(\sum_{m=1}^{M+1}\varphi_m N_m\right)^2\right] dx - \\ - \left[q\sum_{m=1}^{M+1}\varphi_m N_m - \frac{\alpha}{2}\left(\sum_{m=1}^{M+1}\varphi_m N_m - \beta\right)^2\right]. \tag{8.142}$$

Необходимыми условиями минимума функции $I(\varphi_1, \ldots$ $\ldots, \varphi_{M+1})$ многих переменных является равенство нулю ее частных производных по переменным φ_l, $l = \overline{1,\ M+1}$:

$$\frac{\partial I}{\partial \varphi_l} = 0, \quad l = \overline{1,\ M+1}. \tag{8.143}$$

В результате получаем следующую систему $M+1$ линейных алгебраических уравнений относительно φ_l, $l = \overline{1,\ M+1}$:

$$\int_0^1 \left(\frac{dN_l}{dx}\sum_{m=1}^{M+1}\varphi_m \frac{dN_m}{dx} + N_l\sum_{m=1}^{M+1}\varphi_m N_m\right) dx - qN_l + \\ + \alpha N_l\left(\sum_{m=1}^{M+1}\varphi_m N_m - \beta\right) = 0, \quad l = \overline{1, M+1}. \tag{8.144}$$

В силу того что базисные функции N_m, $m = \overline{1,\ M+1}$, в виде равенств (8.3) кусочно-линейны, выражения (8.144) не равны нулю на отрезках расчетной области $x \in [0;\ 1]$, где не равна нулю соответствующая базисная функция.

Поэтому, разбив область [0; 1] на M одномерных элементов и использовав аддитивное свойство определенного интеграла, будем иметь

$$\sum_{e=1}^{M} \int\limits_{x_i^e}^{x_j^e} \left(\frac{dN_l^e}{dx} \sum_{m=1}^{M+1} \varphi_m^e \frac{dN_m^e}{dx} + N_l^e \sum_{m=1}^{M+1} \varphi_m^e N_m^e \right) dx - qN_1 +$$

$$+ \alpha N_{M+1} \left(\sum_{m=1}^{M+1} \varphi_m N_m - \beta \right) = 0, \quad l = \overline{1,\ M+1}, \quad (8.145)$$

где x_i^e, x_j^e — соответственно координаты левого и правого нумерованных узлов элемента Ω^e, причем для внутренних нумерованных узлов третье и четвертое слагаемые равны нулю, для узла $l = 1$ третье слагаемое равно $-q$, а четвертое — нулю и, наконец, для $l = M + 1$ третье слагаемое равно нулю, а четвертое — выражению $\big(\alpha\varphi_{M+1} - \alpha\beta\big)$.

Для конечного элемента с локальными номерами i и j все интегралы под знаком суммы в (8.145) будут равны нулю для элементов, не ассоциируемых с узлами i и j, т. е. для системы алгебраических уравнений

$$\begin{cases} K\varphi = f, \\ k_{lm} = \sum\limits_{e=1}^{M} k_{lm}^e, \\ f_l = \sum\limits_{e=1}^{M} f_l^e, \end{cases} \quad (8.146)$$

где $k_{lm}^e = 0$, если $l, m \neq i, j$;

$$k_{ij}^e = k_{ji}^e = \int\limits_{x_i^e}^{x_j^e} \left(\frac{dN_i^e}{dx} \frac{dN_j^e}{dx} + N_i^e N_j^e \right) dx,$$

если $l = i$, $m = j$ или $l = j$, $m = i$;

$$k_{ii}^e = k_{jj}^e = \int\limits_{x_i}^{x_j} \left[\left(\frac{dN_i^e}{dx} \right)^2 + (N_i^e)^2 \right] dx,$$

если $l = m = i$ или $l = m = j$.

Таким образом, в каждый элемент k_{lm} матрицы K и компонент f_l вектора f дают вклад только те конечные элементы, которые ассоциируются с глобальными номерами узлов l и m.

УПРАЖНЕНИЯ

1. Используя подходящую систему базисных функций в виде многочленов, аппроксимировать функцию $\varphi = 1 + \sin(\pi x/2)$ на отрезке $[0;\ 1]$. Применить методы коллокаций и Галеркина.

2. Для одномерной стационарной задачи теплопроводности

$$d^2\varphi/dx^2 + \varphi + 1 = 0;$$

$$\varphi(0) = 0;$$

$$d\varphi(1)/dx = -\varphi$$

найти приближенное решение методом Галеркина, исследовать скорость сходимости путем сравнения с аналитическим решением.

3. Рассмотреть функционал

$$I(\varphi) = \int\limits_{\Omega} \left[\frac{\lambda}{2} \left(\frac{\partial\varphi}{\partial x} \right)^2 + \frac{\lambda}{2} \left(\frac{\partial\varphi}{\partial y} \right)^2 - Q\varphi \right] d\Omega - \\ - \int\limits_{\Gamma} \left[\alpha/2 \cdot \varphi^2 - q\varphi \right] d\Gamma,$$

где λ, Q, φ суть функции только x, y. Найти уравнение Эйлера и вид естественных краевых условий.

4. Доказать, что следующие операторы являются симметричными и положительно определенными:

а) $L = -d^2/dx^2$ относительно функций, удовлетворяющих условиям

$$d\varphi(0)/dx + a\varphi(0) = 0;$$

$$d\varphi(1)/dx + b\varphi(1) = 0;$$

$$a < 0;\quad b > 0;$$

б) $L\varphi = d^2\left[a(x)\, d^2\varphi/dx^2\right]/dx^2 + b(x)\,\varphi$ относительно множества функций, удовлетворяющих условиям $\varphi = d\varphi/dx = 0$ на $x = 0$ и на $x = 1$;

в) оператор $\partial^4/\partial x^4 + 2\partial^4/\partial x^2\partial y^2 + \partial^4/\partial y^4$ в области Ω, ограниченной замкнутой границей Γ относительно функций, удовлетворяющих условию $\varphi = \partial\varphi/\partial n = 0$ на Γ.

5. Решить задачи конечно-элементным вариационным методом:

$$\text{а) } d^2\varphi/dx^2 + \exp(-x) = 0;\ \varphi(0) = 0;\ d\varphi(1)/dx = 1;$$

$$\text{б) } d^4\varphi/dx^4 = \sin\pi x;\ \varphi(0) = \varphi(1) = 0;\ \frac{d\varphi(0)}{dx} = \frac{d\varphi(1)}{dx} = 0.$$

6. Конечно-элементным методом Галеркина решить задачи в квадрате $x \in [0;\,1]$, $y \in [0;\,1]$, разбив его на 18 треугольных конечных элементов:

а) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0,$$

$$u(x,0) = x,$$

$$u(x,1) = x^2 + 1,$$

$$u(0,y) = y,$$

$$u(1,y) = y^2 + 1;$$

б) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0,$$

$$u(x,0) = x,$$

$$u(x,1) = x^2 + 1,$$

$$u(0,y) = y,$$

$$\frac{u(1,0)}{\partial x} + u(1,0) = 1;$$

в) $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2},$$

$$u(x, 0, t) = x,$$

$$u(x, 1, t) = x^2 + 1,$$

$$u(0, y, t) = y,$$

$$u(1, y, t) = y^2 + 1;$$

$$u(x, y, 0) = 1.$$

ГЛАВА IX

МЕТОД ГРАНИЧНЫХ ЭЛЕМЕНТОВ РЕШЕНИЯ МНОГОМЕРНЫХ СТАЦИОНАРНЫХ ЗАДАЧ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Программа

Метод граничных элементов (ГЭ) решения задач математической физики. Использование фундаментальных решений, основной интегральной формулы Грина и граничного интегрального уравнения. Формирование ГЭ, базисных функций и итоговой СЛАУ. Достоинства и недостатки метода ГЭ.

Метод граничных элементов основан на представлении краевой задачи для уравнений Лапласа или Пуассона в виде граничного интегрального уравнения с использованием во второй формуле Грина фундаментальных (без краевых условий) решений уравнения Лапласа [30].

Известно, что фундаментальными решениями уравнения Лапласа $\Delta u = 0$ являются функции

$$u(M) = \ln \frac{1}{R(M, M_W)} = -\ln \sqrt{(x_M - x_W)^2 + (y_M - y_W)^2} \tag{9.1}$$

в двумерном случае и

$$u(M) = \frac{1}{R(M, M_W)} = \left(\sqrt{(x_M - x_W)^2 + (y_M - y_W)^2 + (z_M - z_W)^2}\right)^{-1} \tag{9.2}$$

в трехмерном случае. В (9.1), (9.2) $R(M, M_W)$ — расстояние между точкой расчетной области $M \in \Omega$ и граничной точкой $M_W \in \Gamma$.

Применяя к фундаментальному решению (9.2) и к искомой функции u вторую формулу Грина, получаем в трехмерном случае тождество, называемое *основной интегральной формулой Грина*:

$$u(M) = \frac{1}{4\pi} \iint\limits_{\Gamma} \left[\frac{1}{R(M, M_W)} \cdot \frac{\partial\, u(M_W)}{\partial\, n} - \right.$$

$$\left. -u(M_W) \frac{\partial}{\partial\, n} \left(\frac{1}{R(M, M_W)} \right) \right] d\,\Gamma(M_W) - \iiint\limits_{\Omega} \frac{\Delta u(M')}{R(M, M')}\, d\Omega, \qquad (9.3)$$

где n — направление внешней нормали к поверхности Γ (для внутренней нормали знак в тождестве меняется на противоположный). Если u удовлетворяет уравнению Лапласа, то объемный интеграл в (9.3) равен нулю, а для уравнения Пуассона этот интеграл известен и в дальнейшем это слагаемое рассматриваться не будет. Здесь $M' \in \Omega$ $(M \neq M')$ — переменная точка в области Ω.

Для произвольной граничной точки $P_W \in \Gamma$ в (9.3) необходимо перейти к пределу при стремлении внутренней точки $M \in \Omega$ к точке $P_W \in \Gamma$. Получим следующее *интегральное уравнение*:

$$u(P_W) = \frac{1}{2\pi} \iint\limits_{\Gamma} \left[\frac{1}{R(P_W, M_W)} \frac{\partial\, u(M_W)}{\partial n} - \right.$$

$$\left. - u(M_W) \frac{\partial}{\partial\, n} \left(\frac{1}{R(P_W, M_W)} \right) \right] d\,\Gamma, \qquad (9.4)$$

называемое *граничным интегральным уравнением.*

Если задано граничное условие 1-го рода

$$u(M_W) = \varphi(x, y, z), \quad M_W \in \Gamma, \qquad (9.5)$$

то значения $u(P_W)$ и $u(M_W)$ в (9.4) известны, поэтому из (9.4) можно определить значения $\dfrac{\partial u(M_W)}{\partial n}$, которые, будучи подставленными в (9.3), явно определяют значения $u(M)$, $M \in \Omega$, в нужных точках расчетной области.

При задании граничного условия второго рода

$$\frac{\partial\, u(M_W)}{\partial\, n} = \varphi(x, y, z), \quad M_W \in \Gamma, \qquad (9.6)$$

или третьего рода

$$\frac{\partial u(M_W)}{\partial n} + \alpha u(M_W) = \varphi(x, y, z), \quad M_W \in \Gamma, \qquad (9.7)$$

нормальная производная $\dfrac{\partial u(M_W)}{\partial n}$ на границе Γ определяется или явно из (9.6) или через $u(M_W)$ из (9.7). Подставляя затем $\dfrac{\partial u(M_W)}{\partial n}$ в (9.4), решаем полученное интегральное уравнение относительно $u(M_W)$, в результате чего на границе Γ будут известны и $u(M_W)$, и $\dfrac{\partial u(M_W)}{\partial n}$. Подставляя затем их в (9.3), получаем явное выражение для определения $u(M)$, $M \in \Omega$.

В двумерном случае основная интегральная формула Грина и граничное интегральное уравнение имеют соответственно следующий вид:

$$u(M) = \frac{1}{2\pi}\int\limits_{\Gamma}\left[\ln\frac{1}{R(M,M_W)}\frac{\partial u(M_W)}{\partial n} - \right. \\ \left. -u(M_W)\frac{\partial}{\partial n}\left(\ln\frac{1}{R(M,M_W)}\right)\right]d\Gamma(M_W), \quad (9.8)$$

$$u(P_W) = \frac{1}{\pi}\int\limits_{\Gamma}\left[\ln\frac{1}{R(P_W,M_W)}\frac{\partial u(M_W)}{\partial n} - \right. \\ \left. -u(M_W)\frac{\partial}{\partial n}\left(\ln\frac{1}{R(P_W,M_W)}\right)\right]d\,\Gamma(M_W). \quad (9.9)$$

Таким образом, достоинством *метода граничных интегральных уравнений* является снижение на единицу порядка дифференциального уравнения. К недостаткам можно отнести необходимость отыскания фундаментальных решений, которые известны не для всех уравнений математической физики.

В соответствии с *методом граничных элементов* для численного решения граничного интегрального уравнения (9.9) в двумерном случае граница Γ разбивается точками $P_{W_K} \in \Gamma$, $k = \overline{1,n}$, на n частей (граничных элементов $\Delta\Gamma^k$), причем узлы с номерами 1 и $n+1$ совпадают. Искомая функция $u(P_W)$ аппроксимируется следующей линейной комбинацией базисных функций $N_k(\varphi)$:

$$u \approx \widetilde{u} = \sum_{k=1}^{n} u_k N_k(\varphi), \quad (9.10)$$

где коэффициенты u_k линейной комбинации суть узловые значения искомой функции на границе Γ, а φ — полярный угол радиуса-вектора, соединяющего каждую точку $P_{W_i} \in \Gamma$, $i = \overline{1,n}$,

как центра локальной полярной системы координат, с точками M_{W_k}, $k = \overline{1, n}$, $k \neq i$.

В качестве *базисных функций* N_k, $k = \overline{1, n}$, рассматриваются кусочно-линейные функции

$$N_k = \begin{cases} (\varphi - \varphi_{k-1})/(\varphi_k - \varphi_{k-1}), & \varphi_{k-1} \leqslant \varphi \leqslant \varphi_k; \\ (\varphi_{k+1} - \varphi)/(\varphi_{k+1} - \varphi_k), & \varphi_k \leqslant \varphi \leqslant \varphi_{k+1}; \\ 0, \quad \varphi < \varphi_{k-1}, \quad \varphi > \varphi_{k+1}, \end{cases} \tag{9.11}$$

если за нумерованные узлы k принять точки разбиения, или кусочно-постоянные функции

$$N_k = \begin{cases} 1, & \varphi_{k-1} \leqslant \varphi \leqslant \varphi_k; \\ 0, & \varphi < \varphi_{k-1}, \quad \varphi > \varphi_k, \end{cases} \tag{9.12}$$

если за нумерованные узлы принять точки P_k, лежащие в геометрических центрах граничных элементов $\Delta\Gamma^k$. Именно последний вид базисных функций, как более простых, будет использован в данной главе.

Представляя криволинейный интеграл в граничном интегральном уравнении (9.9) в виде суммы по граничным элементам:

$$\pi\, u(P_{W_i}) - \sum_{k=1,\, k\neq i}^{n} \int\limits_{\Delta\Gamma^k} \left\{ \widetilde{u}(M_{W_k}) \frac{\partial}{\partial n} \left[\ln R(P_{W_i}, M_{W_k})\right] - \right.$$

$$\left. - \ln R(P_{W_i}, M_{W_k}) \frac{\partial\, \widetilde{u}(M_{W_k})}{\partial n} \right\} d\,\Gamma(M_W) = 0, \quad i = \overline{1, n}, \tag{9.13}$$

и подставляя в (9.13) вместо $\widetilde{u}$ линейную комбинацию (9.10), получаем систему линейных алгебраических уравнений относительно u_{W_k}, $k = \overline{1, n}$.

Рассмотрим применение метода граничных элементов для численного решения стационарной задачи теплопроводности в многосвязных областях, которая формулируется в виде следующей краевой задачи для квазилинейного уравнения теплопроводности (рис. 9.1):

$$\frac{\partial}{\partial x}\left(\lambda(u)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda(u)\frac{\partial u}{\partial y}\right) = 0, \tag{9.14}$$

$$\alpha_{W_1}(u_{e1} - u_{W_1}) - \lambda(u)\left.\frac{\partial u}{\partial n}\right|_{W_1} = 0, \tag{9.15}$$

$$\alpha_{W_2}(u_{e2} - u_{W_2}) + \lambda(u)\left.\frac{\partial u}{\partial n}\right|_{W_2} = 0. \tag{9.16}$$

Поскольку использование для решения интегральной формулы Грина (9.8) и граничного интегрального уравнения (9.9) пред-

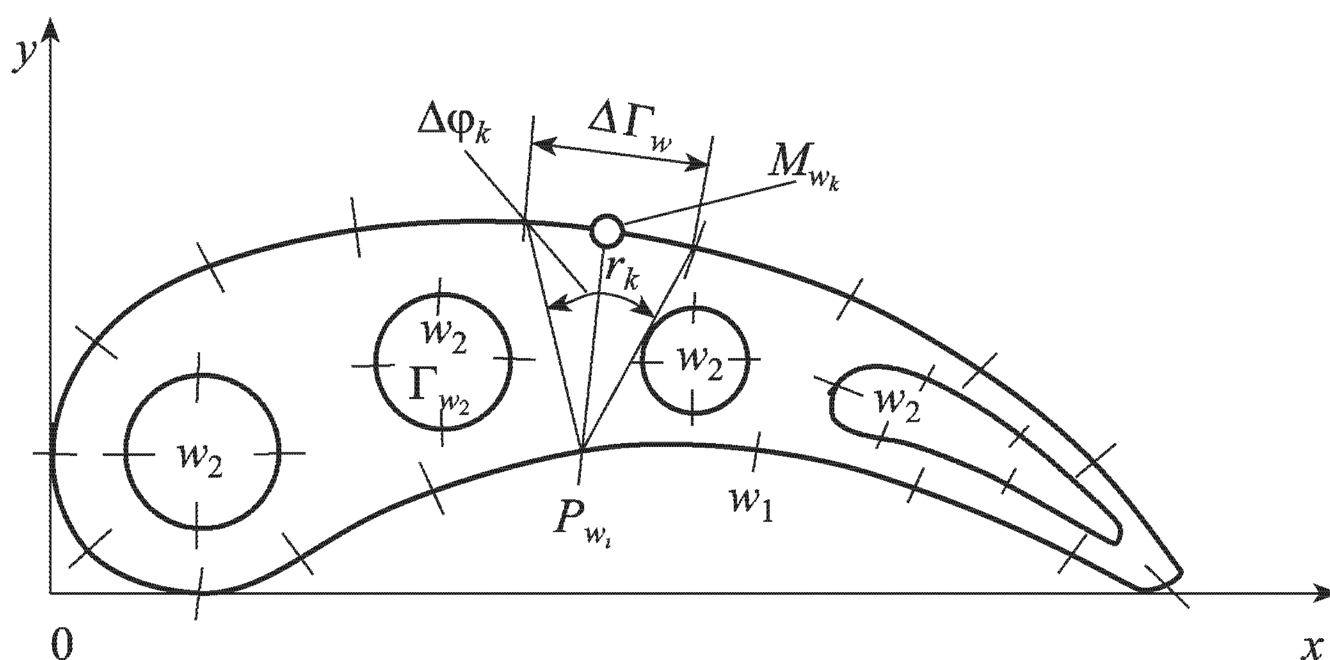


Рис. 9.1. Разбиение границы многосвязной области

полагает линейность задачи, необходимо задачу (9.14)–(9.16) линеаризовать, для чего используется следующая замена искомой переменной, называемая подстановкой Кирхгофа:

$$\Lambda = \int_0^u \lambda(\xi)d\xi. \tag{9.17}$$

Тогда

$$\frac{\partial}{\partial x}\left(\lambda(u)\frac{\partial u}{\partial x}\right) = \frac{\partial}{\partial x}\left[\lambda(u)\frac{\partial u}{\partial \Lambda}\frac{\partial \Lambda}{\partial x}\right] =$$

$$= \frac{\partial}{\partial x}\left[\lambda(u)\frac{du}{d\Lambda}\frac{\partial \Lambda}{\partial x}\right] = \frac{\partial}{\partial x}\left[\lambda(u)\left(\frac{d\Lambda}{du}\right)^{-1}\frac{\partial \Lambda}{\partial x}\right] = \frac{\partial^2 \Lambda}{\partial x^2};$$

$$\frac{\partial}{\partial y}\left(\lambda(u)\frac{\partial u}{\partial y}\right) = \frac{\partial^2 \Lambda}{\partial y^2},$$

и уравнение (9.14) трансформируется в следующее уравнение Лапласа:

$$\frac{\partial^2 \Lambda}{\partial x^2} + \frac{\partial^2 \Lambda}{\partial y^2} = 0. \tag{9.18}$$

В краевых условиях (9.15), (9.16) подстановка (9.17) линеаризует слагаемое с нормальной производной искомой функции, однако линейные члены становятся нелинейными. Для сохранения линейности этих слагаемых примем в начальном приближении $\lambda(u)$ постоянным и равным λ_c. Тогда из (9.17) имеем

$$u = \frac{\Lambda}{\lambda_c}, \tag{9.19}$$

и краевые условия (9.15), (9.16) будут трансформированы следующим образом:

$$\alpha_{W_1}(u_{e1} - \Lambda_{W_1}/\lambda_c) - \left.\frac{\partial \Lambda}{\partial n}\right|_{W_1} = 0, \tag{9.20}$$

$$\alpha_{W_2}(u_{e2} - \Lambda_{W_2}/\lambda_c) + \left.\frac{\partial \Lambda}{\partial n}\right|_{W_2} = 0. \tag{9.21}$$

Если решение $\Lambda(x, y)$ в точке (x, y) линейной третьей краевой задачи (9.18), (9.20), (9.21) для уравнения Лапласа подставить в (9.17) и после интегрирования решить соответствующее алгебраическое уравнение, степень которого на единицу выше степени функции $\lambda(u)$, то получим значение искомой функции $u(x, y)$ в той же точке. При этом в радикалах решается уравнение степени не выше четвертой:

$$a_0 u^4 + a_1 u^3 + a_2 u^2 + a_3 u + a_4 = \Lambda, \tag{9.22}$$

что соответствует заданию $\lambda(u)$ в виде многочлена степени не выше третьей.

Итак, решается стационарная задача (9.18), (9.20), (9.21) методом граничных элементов с использованием базисных функций (9.12) в многосвязных областях, представленных на рис. 9.1.

Всякая функция $\Lambda(x, y)$, непрерывная вместе с первыми производными в замкнутой области $\Omega + \Gamma_W$, где $\Gamma_W = \Gamma_{W_1} + \Gamma_{W_2}$ — достаточно гладкая граница, и имеющая вторые производные внутри Ω, удовлетворяет основной интегральной формуле Грина

(9.8) и граничному интегральному уравнению (9.9), которые для задачи (9.18), (9.20), (9.21) имеют вид

$$\Lambda(M) = \frac{1}{2\pi}\int\limits_{\Gamma}\left[\Lambda(M_W)\frac{\partial}{\partial n}(\ln R(M, M_W)) - \right.$$
$$\left. - \ln R(M, M_W)\frac{\partial\Lambda(M_W)}{\partial n}\right] d\Gamma(M_W), \quad (9.23)$$

$$\Lambda(P_W) = \frac{1}{\pi}\int\limits_{\Gamma}\left[\Lambda(M_W)\frac{\partial}{\partial n}\ln R(P_W, M_W) - \right.$$
$$\left. - \ln R(P_W, M_W)\frac{\partial\Lambda(M_W)}{\partial n}\right] d\Gamma(M_W). \quad (9.24)$$

Упростим подынтегральные выражения в (9.23), (9.24), для чего введем локальную полярную систему координат с центром

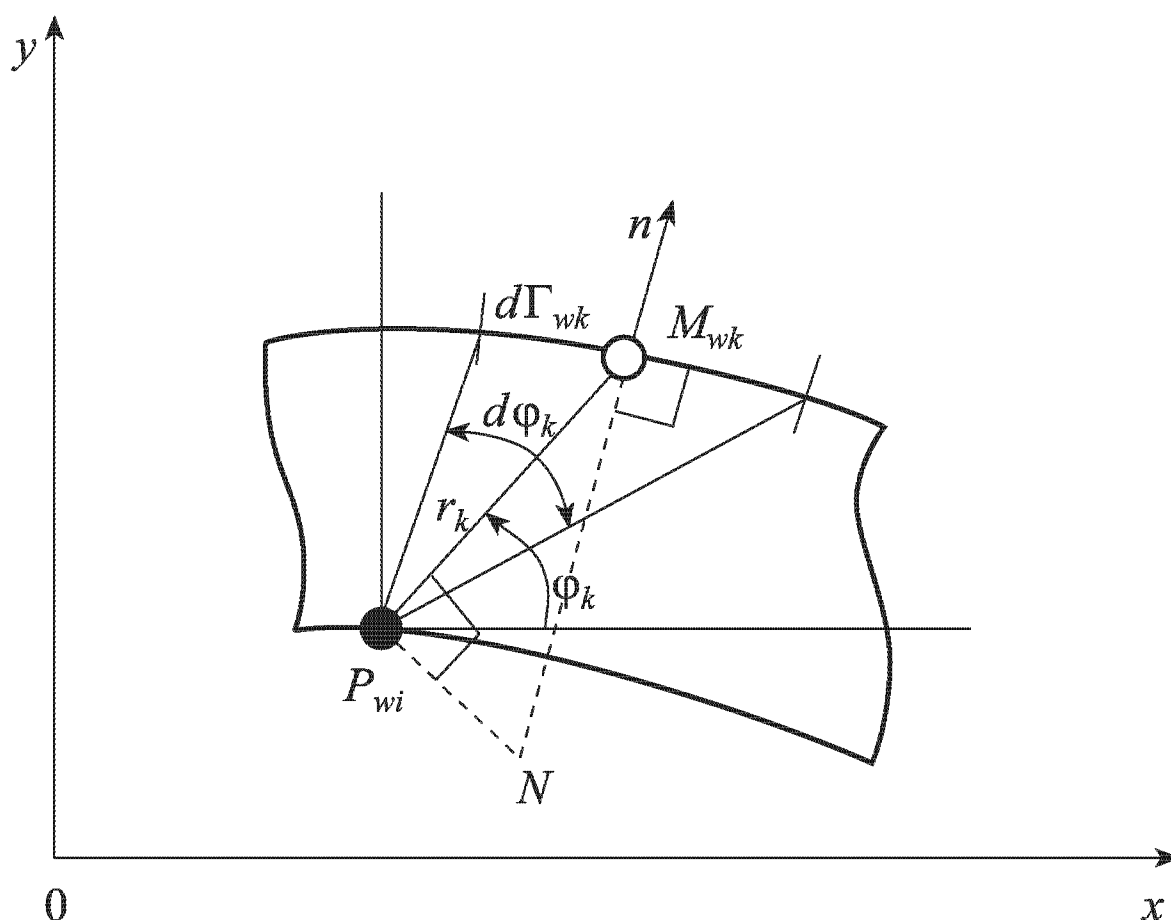


Рис. 9.2. Локальная полярная система координат

в граничной точке P_W, полярным расстоянием $r \equiv R(P_W, M_W)$ и углом φ (рис. 9.2). Тогда в (9.24) первое слагаемое подынтегрального выражения преобразуется следующим образом ($\Gamma_W \equiv \equiv \Gamma(M_W)$):

$$\Lambda(M_W)\frac{\partial}{\partial n}(\ln r)d\Gamma_W = \Lambda(M_W)\frac{1}{r}\frac{\partial r}{\partial n}d\Gamma_W =$$

$$= \Lambda(M_W)\frac{1}{r}\cos(\overline{r},\overline{n})\sqrt{r^2+(dr/d\varphi)^2}\,d\varphi =$$

$$= \Lambda(M_W)\frac{r}{r}\frac{\sqrt{r^2+(dr/d\varphi)^2}}{\sqrt{r^2+(dr/d\varphi)^2}}d\varphi = \Lambda(M_W)d\varphi, \quad (9.25)$$

так как из равенств

$$x = r\cdot\cos\varphi, \quad y = r\cdot\sin\varphi, \quad d\Gamma_W = \sqrt{dx^2+dy^2}$$

следует, что

$$dx = \frac{\partial\, x}{\partial\, r}dr + \frac{\partial\, x}{\partial\varphi}d\varphi = \cos\varphi\, dr - r\cdot\sin\varphi\, d\varphi;$$

$$dy = \frac{\partial\, y}{\partial\, r}dr + \frac{\partial\, y}{\partial\,\varphi}d\varphi = \sin\varphi\, dr + r\cdot\cos\varphi\, d\varphi;$$

$$d\Gamma_W = \sqrt{dr^2+r^2d\varphi^2} = \sqrt{r^2+(dr/d\varphi)^2}\,d\varphi. \quad (9.26)$$

Кроме этого, если элемент касательной к дуге в точке M_W определяется выражением (9.26), то отрезок нормали $M_W N$ имеет длину

$$|M_W N| = \sqrt{r^2+(dr/d\varphi)^2}$$

и, следовательно,

$$\frac{\partial r}{\partial n} = \cos(\overline{r},\overline{n}) = \frac{r}{|M_W N|} = \frac{r}{\sqrt{r^2+(dr/d\varphi)^2}}. \quad (9.27)$$

Таким образом, из (9.26), (9.27) следует (9.25). Используя выражение (9.25) и краевые условия (9.20), (9.21) в граничном интегральном уравнении (9.24), получаем это уравнение для граничных точек P_W в следующем виде:

$$\Lambda(P_W) = \frac{1}{\pi}\int\limits_{\Gamma_W}\Lambda(M_W)d\varphi - \frac{1}{\pi}\int\limits_{\Gamma_W}\ln r\frac{\alpha_W}{\lambda_c}\,[\lambda_c u_e - \Lambda(M_W)]d\Gamma_W, \quad (9.28)$$

где

$$\alpha_W = \left\{\alpha_{W_1},\alpha_{W_2}\right\}, \quad u_e = \{u_{e1},u_{e2}\}, \quad \Gamma_W = \{\Gamma_{W_1},\Gamma_{W_2}\}.$$

Разобьем границу $\Gamma_W = \Gamma_{W_1}+\Gamma_{W_2}$ многосвязной расчетной области на n граничных элементов $\Delta\Gamma_W^i$, $i=\overline{1,n}$, и будем вычислять с помощью формулы (9.28) функцию Λ в нумерованных

узлах P_{W_i}, $i = \overline{1, n}$, находящихся в середине граничных элементов $\Delta\Gamma_W^i$. Количество граничных элементов на наружной границе W_1 принимаем равным нескольким десяткам, уменьшая шаг разбиения в местах увеличения кривизны наружной границы. На каждой из внутренних границ W_2 принимается до десяти граничных элементов, но не меньше четырех.

Используя аддитивное свойство криволинейного интеграла, заменяем интегралы в (9.28) суммами интегралов по граничным элементам $\Delta\Gamma_W^i$ и считаем величины α_W, u_e, $\Lambda(P_W)$ постоянными в пределах каждого граничного элемента. Получаем следующую систему n алгебраических уравнений относительно значений функции $\Lambda(P_{W_i}) \equiv \Lambda_{W_i}$, $i = \overline{1, n}$:

$$- \pi\Lambda_{W_i} + \left(\sum_{k=1}^{i-1, i\neq 1} + \sum_{k=i+1, i\neq n}^{n} \right) (a_k \Lambda_{W_k}) =$$

$$= \left(\sum_{k=1}^{i-1, i\neq 1} + \sum_{k=i+1, i\neq n}^{n} \right) (b_k), \quad i = \overline{1, n}, \qquad (9.29)$$

где

$$a_k = \int\limits_{\varphi_{k-1/2}}^{\varphi_{k+1/2}} d\varphi + \frac{\alpha_{W_k}}{\lambda_c} \int\limits_{\Delta\Gamma_W^k} \ln r \cdot d\Gamma_W,$$

$$b_k = \alpha_{W_k} u_{ek} \int\limits_{\Delta\Gamma_W^k} \ln r \cdot d\Gamma_W.$$

В выражениях (9.29) для каждой зафиксированной точки i-го участка ($i = \overline{1, n}$) пробегаются все k ($k = \overline{1, i-1}$, $i \neq 1$ и $k = \overline{i+1, n}$, $i \neq n$) срединных точек M_{W_k}, за исключением точки P_{W_i}. При этом угол φ изменяется от 0 до π.

Вычисление интегралов в коэффициентах a_k, b_k можно осуществить с помощью квадратурных формул трапеций или Симпсона.

Система (9.29) является системой с наполненной матрицей (т. е. очень мало нулевых элементов матрицы системы). Для ее решения можно использовать как прямые, так и итерационные методы.

Результатом решения системы (9.29) будет распределение функции Λ_{W_i}, $i = \overline{1, n}$, на границе Γ_W расчетной области.

Гранично-элементная аппроксимация основной интегральной формулы Грина (9.23) будет аналогична аппроксимации (9.29) с той лишь разницей, что вместо коэффициента $1/\pi$ при криволинейном интеграле будет стоять коэффициент $1/2\pi$, а значения функций $\Lambda(M_W)$, $\partial\Lambda(M_W)/\partial n$ уже известны из решения системы алгебраических уравнений (9.29) и краевых условий (9.20), (9.21). Таким образом, получаем явные выражения для определения $\Lambda(M)$, $M \in \Omega$, в следующем виде:

$$\Lambda(M) = \frac{1}{2\pi}\sum_{k=1}^{n}(a_k\Lambda_{W_k}) - \frac{1}{2\pi}\sum_{k=1}^{n} b_k, \quad M \in \Omega. \tag{9.30}$$

Здесь в коэффициентах a_k, b_k углы φ — это углы между линиями MM_{W_k}, соединяющими точку $M \in \Omega$ с точками $M_{W_k} \in \Gamma_W$, $k = \overline{1, n}$, и осью x, которые изменяются от 0 до 2π.

После определения функции Λ в области $\Omega + \Gamma_W$ из подстановки (9.17) определяется искомая функция $u(x, y)$ в каждой точке расчетной области. Например, если $\lambda(u)$ задана в виде многочлена третьей степени

$$\lambda(u) = a_0u^3 + a_1u^2 + a_2u + a_3,$$

то в каждой точке расчетной области необходимо решить алгебраическое уравнение четвертой степени

$$\frac{a_0}{4}u^4 + \frac{a_1}{3}u^3 + \frac{a_2}{2}u^2 + a_3u = \Lambda(M), \quad M \in \Omega + \Gamma,$$

которое можно решить в радикалах [30].

В результате распределение функции $u\,(x, y)$ в области $\Omega + \Gamma$ будет определяться в первом приближении, поскольку граничные условия (9.20), (9.21) учитывали постоянное значение коэффициента λ_c. В соответствии с этим будем обозначать это решение верхним индексом (1) $u^{(1)}$: (соответственно $\Lambda^{(1)}$).

Для определения второго приближения $\Lambda^{(2)}$ (соответственно $u^{(2)}$) разложим функцию $\Lambda^{(2)}(u)$ в ряд Тейлора в окрестности $u^{(1)}$, оставив в нем только линейные члены:

$$\Lambda^{(2)}(u) = \Lambda^{(1)} + \frac{d\Lambda^{(1)}}{du}\Delta u + O(\Delta u^2). \tag{9.31}$$

Полагая здесь

$$\Delta u = u^{(2)} - u^{(1)} = \left(\Lambda^{(2)} - \Lambda^{(1)}\right) \Big/ \frac{d\Lambda^{(1)}}{du},$$

получаем

$$u^{(2)} = u^{(1)} - \frac{\Lambda^{(1)}}{\lambda(u^{(1)})} + \frac{\Lambda^{(2)}}{\lambda(u^{(1)})}. \tag{9.32}$$

Применим теперь к производным исходной задачи (9.14)–(9.16) подстановку Кирхгофа (9.17), а вместо искомой функции в краевых условиях (9.15), (9.16) — приближение (9.32), получим следующую третью краевую задачу для уравнения Лапласа относительно функции $\Lambda^{(2)}$:

$$\frac{\partial^2 \Lambda^{(2)}}{\partial x^2} + \frac{\partial^2 \Lambda^{(2)}}{\partial y^2} = 0, \quad (x, y) \in \Omega; \tag{9.33}$$

$$\alpha_{W_1}^{(2)} \left(u_{e1}^{(2)} - \Lambda_{W_1}^{(2)}\right) = \left.\frac{\partial \Lambda^{(2)}}{\partial n}\right|_{W_1}, \quad (x, y) \in \Gamma_{W_1}; \tag{9.34}$$

$$\alpha_{W_2}^{(2)} \left(u_{e2}^{(2)} - \Lambda_{W_2}^{(2)}\right) = -\left.\frac{\partial \Lambda^{(2)}}{\partial n}\right|_{W_2}, \quad (x, y) \in \Gamma_{W_2}, \tag{9.35}$$

где

$$\alpha_{W_1}^{(2)} = \alpha_{W_1} / \lambda(u_{W_1}^{(1)}); \quad \alpha_{W_2}^{(2)} = \alpha_{W_2} / \lambda(u_{W_2}^{(1)});$$

$$u_{e1}^{(2)} = \lambda(u_{W_1}^{(1)})(u_{e1} - u_{W_1}^{(1)}) + \Lambda_{W_1}^{(1)};$$

$$u_{e2}^{(2)} = \lambda(u_{W_2}^{(1)})(u_{e2} - u_{W_2}^{(1)}) + \Lambda_{W_2}^{(1)}.$$

Из сравнения задачи (9.33)–(9.35) с задачей (9.18), (9.20), (9.21) следует, что для определения $\Lambda^{(2)}$ формально можно использовать решение (9.29) для точек на границе Γ_W и решение (9.30) для внутренних точек области Ω, заменяя коэффициенты a_k, b_k следующим образом:

$$a_k = \int\limits_{\varphi_{k-1/2}}^{\varphi_{k+1/2}} d\varphi + \alpha_{W_k}^{(2)} \int\limits_{\Delta\Gamma_W^k} \ln r \cdot d\Gamma_W,$$

$$b_k = \alpha_{W_k}^{(2)} u_{ek}^{(2)} \int\limits_{\Delta\Gamma_W^k} \ln r \cdot d\Gamma_W,$$

где

$$\alpha_W^{(2)} = \left\{\alpha_{W_1}^{(2)}, \alpha_{W_2}^{(2)}\right\}; \quad u_e^{(2)} = \left\{u_{e1}^{(2)}, u_{e2}^{(2)}\right\}; \quad \Gamma_W = \{\Gamma_{W_1}, \Gamma_{W_2}\}.$$

Искомая функция $u^{(2)}$ в точках $M \in \Omega + \Gamma$ определяется путем решения уравнения (9.17).

СПИСОК ЛИТЕРАТУРЫ

1. *Демидович Б. П.*, *Марон И. А.* Основы вычислительной математики. — М.: ГИФМЛ. 1963.
2. *Самарский А. А.*, *Николаев Е. С.* Методы решения разностных уравнений. — М.: Наука. 1978.
3. *Бахвалов Н. С.* Численные методы. — М.: Наука. 1975.
4. *Форсайт Дж.*, *Молер К.* Численные методы решения систем линейных алгебраических уравнений. — М.: Мир, 1969.
5. *Крылов В. И.*, *Бобков В. В.*, *Монастырный П. И.* Вычислительные методы. — М.: Наука. 1976. Т. 1, 2.
6. *Плис А. И.*, *Сливина Н. А.* Лабораторный практикум по высшей математике. — М.: Высшая школа. 1983.
7. *Турчак Л. И.* Основы численных методов. — М.: Наука. 1980.
8. *Корн Г.*, *Корн Т.* Справочник по математике для научных работников и инженеров. — М.: Наука. 1968.
9. Сборник задач по математике для ВТУЗов. Методы оптимизации, уравнения в частных производных, интегральные уравнения/ *Под. ред. Ефимова А. В.* — М.: Наука. 1990.
10. *Самарский А. А.*, *Гулин А. В.* Устойчивость разностных схем. — М.: Наука. 1973.
11. *Самарский А. А.* Теория разностных схем. — М.: Наука. 1983.
12. *Самарский А. А.*, *Попов Ю. П.* Разностные методы решения задач газовой динамики. — М.: Наука. 1980.
13. *Самарский А. А.*, *Гулин А. В.* Численные методы. — М.: Наука. 1989.
14. *Яненко Н. Н.* Метод дробных шагов решения многомерных задач математической физики. — Новосибирск: Наука. 1967.

15. *Саульев В. И.* Интегрирование уравнений параболического типа методом сеток. — М.: Физматгиз. 1960.

16. *Марчук Г. И.* Методы расщепления. — М.: Наука. 1988.

17. *Формалев В. Ф.* Экономичный абсолютно устойчивый алгоритм численного решения двумерных уравнений параболического типа. //Сборник научных трудов «Численные методы решения задач аэродинамики». — М: Издательство МАИ. 1987. С. 56–59.

18. *Формалев В. Ф.*, *Тюкин О. А.* Экономичный абсолютно устойчивый метод расщепления с экстраполяцией численного решения задач, содержащих смешанные дифференциальные операторы. //Вычислительные технологии. 1995. Т. 4. № 10. С. 290–299.

19. *Формалев В. Ф.*, *Воробьева О. Р.* Метод переменных направлений с экстраполяцией численного решения задач теплопроводности с тензором теплопроводности и конвективными членами. — М: Вестник МАИ. 1997. Т. 5. № 1. С. 41–48.

20. *Формалев В. Ф.*, *Тюкин О. А.* Неявный экономичный метод численного решения параболических задач, содержащих смешанные производные //Математическое моделирование.1996. Т. 8. № 6. С. 27–32.

21. *Формалев В. Ф.*, *Тюкин О. А.* Неявный метод дробных шагов с расщеплением смешанных дифференциальных операторов.// Вычислительные технологии. 1998. Т. 3. № 6. С. 82–91.

22. *Березин И. С.*, *Жидков Н. П.* Методы вычислений. — М.: Наука. 1960. Т. 2.

23. *Котляр Я. М.* Методы математической физики в прикладных задачах. — М.: Издательство МАИ. 1978.

24. *Пирумов У. Г.* Численные методы. — М.: Издательство МАИ.

25. *Годунов С. К.*, *Забродин А. В.*, *Иванов Н. Е. и др.* Численные методы решения многомерных задач газовой динамики. — М.: Наука. 1976.

26. *Сегерлинд Л.* Применение метода конечных элементов. — М.: Мир. 1979.

27. *Зенкевич О.*, *Морган К.* Конечные элементы и аппроксимация. — М.: Мир. 1986.
28. *Формалев В. Ф.* Метод конечных элементов в задачах теплообмена. — М.: Издательство МАИ. 1991.
29. *Ши Д.* Численные методы в задачах теплообмена. — М.: Мир. 1988.
30. *Галицейский Б. М.*, *Совершенный В. Д.*, *Формалев В. Ф. и др.* Тепловая защита лопаток турбин. — М.: Издательство МАИ. 1996.
31. *Тихонов А. Н.*, *Самарский А. А.* Уравнения математической физики. — М.: Наука. 1972.

ОГЛАВЛЕНИЕ

I. Численные методы алгебры и анализа

Глава 1. **Элементы теории погрешностей** 11

Глава 2. **Численные методы алгебры** 16

§ 2.1. Численные методы решения СЛАУ 17
2.1.1. Метод Гаусса (17). 2.1.2. Метод прогонки (26). 2.1.3. Обоснование метода прогонки (31). 2.1.4. Матричная прогонка (33). 2.1.5. Нормы векторов и матриц (34). 2.1.6. Итерационные методы решения СЛАУ. Метод простых итераций (38). 2.1.7. Метод Зейделя решения СЛАУ (45). 2.1.8. Метод Зейделя для нормальных СЛАУ (47).

§ 2.2. Численные методы решения нелинейных и трансцендентных уравнений 50
2.2.1. Способы отделения корней (51). 2.2.2. Методы уточнения корней (52). 2.2.3. Скорость сходимости. Процедура Эйткена ускорения сходимости (66). 2.2.4. Замечания к методам отделения корней (68).

§ 2.3. Численные методы решения систем нелинейных уравнений 69
2.3.1. Метод простых итераций и метод Зейделя решения систем нелинейных уравнений (70). 2.3.2. Метод Ньютона (72).

§ 2.4. Численные методы решения задач на собственные значения и собственные векторы матриц линейных преобразований . 76
2.4.1. Основные определения и спектральные свойства матриц (76). 2.4.2. Метод вращений Якоби численного решения задач на собственные значения и собственные векторы матриц (80). 2.4.3. Частичная проблема собственных значений и собственных векторов матрицы. Степенной метод (91).

Глава 3. **Теория приближений** 97

§ 3.1. Исчисление конечных разностей 99
§ 3.2. Задача интерполяции 100
3.2.1. Интерполяционный многочлен Лагранжа (101). 3.2.2. Интерполяционный многочлен Ньютона (103). 3.2.3. Погрешность многочленной интерполяции (104). 3.2.4. Интерполяционный многочлен Ньютона, построенный

с помощью разделенных разностей (106). 3.2.5. Сплайн-интерполяция (108).
§ 3.3. Метод наименьших квадратов 118
§ 3.4. Численное дифференцирование 127
3.4.1. Метод Рунге уточнения формул численного дифференцирования (130).
§ 3.5. Численное интегрирование функций 134
3.5.1. Формула прямоугольников численного интегрирования (135). 3.5.2. Численное интегрирование с помощью формулы трапеций (136). 3.5.3. Формула Симпсона численного интегрирования (140). 3.5.4. Процедура Рунге оценки погрешности и уточнения формул численного интегрирования (144).

Глава 4. **Численные методы решения задач для обыкновенных дифференциальных уравнений** 150

§ 4.1. Основные определения и постановка задач Коши для обыкновенных дифференциальных уравнений 150
§ 4.2. Метод Эйлера численного решения задач Коши для ОДУ и систем ОДУ 154
4.2.1. Метод Эйлера для нормальных систем ОДУ (155).
§ 4.3. Метод Эйлера–Коши (Эйлера с пересчетом) 156
4.3.1. Метод Эйлера–Коши для нормальных систем (157).
§ 4.4. Метод Рунге–Кутта 158
4.4.1. Метод Рунге–Кутта для нормальных систем ОДУ (161).
§ 4.5. Выбор шага численного интегрирования задач Коши 162
§ 4.6. Процедура Рунге оценки погрешности и уточнения численного решения задач Коши 163
§ 4.7. Численные методы решения краевых задач для ОДУ 175
4.7.1. Постановка краевых задач для ОДУ (175). 4.7.2. Конечно-разностный метод с использованием метода прогонки решения краевых задач для ОДУ (176). 4.7.3. Конечно-разностная схема со вторым порядком аппроксимации краевых условий, содержащих производные (178). 4.7.4. Метод пристрелки численного решения краевых задач для ОДУ (180). 4.7.5. Метод пристрелки с использованием итерационной процедуры Ньютона (182).

Глава 5. **Численные методы оптимизации** 195

§ 5.1. Классификация численных методов оптимизации 195
§ 5.2. Численные методы безусловной минимизации функций одной переменной. Прямые методы 196

5.2.1. Метод перебора (197). 5.2.2. Метод деления отрезка пополам (198). 5.2.3. Метод золотого сечения (200).

§ 5.3. Методы минимизации, использующие производные. Метод Ньютона 205

§ 5.4. Безусловная минимизация функций многих переменных . . 207

5.4.1. Метод градиентного спуска (207). 5.4.2. Метод наискорейшего спуска (212). 5.4.3. Метод сопряженных направлений (215).

II. Численные методы решения задач для уравнений математической физики

Глава 6. **Метод конечных разностей** 221

§ 6.1. Постановка задач математической физики 222

6.1.1. Постановка задач для уравнений параболического типа (222). 6.1.2. Постановка задач для уравнений гиперболического типа (224). 6.1.3. Постановка задач для уравнений эллиптического типа (226).

§ 6.2. Основные определения и конечно-разностные схемы для различных задач математической физики 228

6.2.1. Основные определения (228). 6.2.2. Конечно-разностная аппроксимация задач для уравнений гиперболического типа (231). 6.2.3. Конечно-разностная аппроксимация задач для уравнений эллиптического типа (233).

§ 6.3. Основные понятия, связанные с конечно-разностной аппроксимацией дифференциальных задач 235

6.3.1. Аппроксимация и порядок аппроксимации (235). 6.3.2. Устойчивость (236). 6.3.3. Сходимость и порядок сходимости (237). 6.3.4. Теорема эквивалентности о связи аппроксимации и устойчивости со сходимостью (238). 6.3.5. Консервативность и корректность (238).

§ 6.4. Анализ порядка аппроксимации разностных схем 239

§ 6.5. Исследование устойчивости конечно-разностных схем 240

6.5.1. Метод гармонического анализа (241). 6.5.2. Исследование устойчивости методом гармонического анализа явной и неявной схем для уравнения теплопроводности (242). 6.5.3. Исследование устойчивости методом гармонического анализа явной и неявной схем для волнового уравнения (243). 6.5.4. Принцип максимума (245). 6.5.5. Спектральный метод исследования устойчивости (246). 6.5.6. Энергетический метод исследования устойчивости конечно-разностных схем (248).

§ 6.6. Конечно-разностный метод решения задач для уравнений параболического типа . . . 254
6.6.1. Однородные и консервативные конечно-разностные схемы для задач теплопроводности с граничными условиями, содержащими производные (254). 6.6.2. Неявно-явная конечно-разностная схема с весами. Схема Кранка–Николсона (260). 6.6.3. Метод прямых (264).
§ 6.7. Метод конечных разностей решения задач для волнового уравнения с граничными условиями, содержащими производные . . . 269
§ 6.8. Метод установления и его обоснование . . . 273

Глава 7. **Метод конечных разностей решения многомерных задач математической физики. Методы расщепления** . . . 279

§ 7.1. Метод матричной прогонки . . . 281
§ 7.2. Метод переменных направлений Писмена–Рэчфорда . . . 284
§ 7.3. Метод дробных шагов Н. Н. Яненко . . . 288
§ 7.4. Метод переменных направлений с экстраполяцией В. Ф. Формалева . . . 290
7.4.1. Аппроксимация (293). 7.4.2. Устойчивость (296).
§ 7.5. Схема метода полного расщепления Формалева–Тюкина . . 298
§ 7.6. Методы расщепления численного решения эллиптических задач . . . 301
§ 7.7. Методы решения задач для уравнений гиперболического типа . . . 301
7.7.1. Метод характеристик решения квазилинейных гиперболических систем (302). 7.7.2. Метод сквозного счета. Задача о распаде произвольного разрыва. Метод С. К. Годунова (307).

Глава 8. **Метод конечных элементов** . . . 316

§ 8.1. Основы МКЭ . . . 317
§ 8.2. Система базисных функций . . . 318
8.2.1. Кусочно-постоянные базисные функции (319). 8.2.2. Линейные кусочно-непрерывные базисные функции (321).
§ 8.3. Методы взвешенных невязок. Весовые функции . . . 322
8.3.1. Метод поточечной коллокации (323). 8.3.2. Метод Галеркина (324). 8.3.3. Метод наименьших квадратов (325).
§ 8.4. Конечно-элементный метод Галеркина решения краевых задач для обыкновенных дифференциальных уравнений . . . 325
8.4.1. Слабая формулировка метода Галеркина (326). 8.4.2. Формирование локальной и глобальной матриц жест-

кости. Ансамблирование элементов (328). 8.4.3. Случай граничных условий, содержащих производные (333).

§ 8.5. Метод конечных элементов в стационарных задачах математической физики 335

8.5.1. Основные этапы решения стационарных задач математической физики методом конечных элементов (335). 8.5.2. Принципы разбиения плоских областей на конечные элементы (337). 8.5.3. Аппроксимация линейными многочленами и базисные функции (339). 8.5.4. Слабая формулировка конечно-элементного метода Галеркина (342). 8.5.5. Ансамблирование элементов и построение глобальной СЛАУ (348).

§ 8.6. Метод конечных элементов в многомерных нестационарных задачах математической физики 350

§ 8.7. Особенности решения пространственных задач математической физики методом конечных элементов 353

§ 8.8. Оценка погрешности метода конечных элементов 356

8.8.1. Погрешность конечно-элементного метода решения задач для обыкновенных дифференциальных уравнений (356). 8.8.2. Погрешность конечно-элементного метода решения задач для уравнений в частных производных (362).

§ 8.9. Вариационный принцип в МКЭ 365

8.9.1. Введение в вариационное исчисление (366). 8.9.2. Конечно-элементный вариационный принцип на основе симметричного дифференциального оператора. Вариационный метод Релея–Ритца (369). 8.9.3. Решение задач с помощью конечно-элементного вариационного принципа (373).

Глава 9. **Метод граничных элементов решения многомерных стационарных задач математической физики** . . 379

Список литературы 391

Учебное издание

ФОРМАЛЕВ Владимир Федорович
РЕВИЗНИКОВ Дмитрий Леонидович

ЧИСЛЕННЫЕ МЕТОДЫ

Редактор *Н.Б. Бартошевич-Жагель*
Оригинал-макет: *А.Л. Жигарев*
Оформление переплета: *А.Ю. Алехина*

ЛР № 071930 от 06.07.99. Подписано в печать 02.06.04.
Формат 60×90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 25. Уч.-изд. л. 25. Тираж 1000 экз. Заказ №

Издательская фирма «Физико-математическая литература»
МАИК «Наука/Интерпериодика»
117997, Москва, ул. Профсоюзная, 90
E-mail: fizmat@maik.ru, fmlsale@maik.ru
http://www.fml.ru

Отпечатано с готовых диапозитивов
в ППП «Типография «Наука»
121099, Москва, Шубинский пер., 6

ISBN 5-9221-0479-9

